# Face Recognition in Challenging Situations

**Eilidh Clare Noyes**

Doctor of Philosophy

University of York

Psychology

May 2016

# Abstract

A great deal of previous research has demonstrated that face recognition is unreliable for unfamiliar faces and reliable for familiar faces. However, such findings typically came from tasks that used 'cooperative' images, where there was no deliberate attempt to alter apparent identity. In applied settings, images are often far more challenging in nature. For example multiple images of the *same identity* may appear to be different identities, due to either *incidental* changes in appearance (such as age or style related change, or differences in images capture) or *deliberate* changes (evading own identity through disguise). At the same time, images of *different identities* may look like the same person, due to either *incidental* changes (natural similarities in appearance), or *deliberate* changes (attempts to impersonate someone else, such as in the case of identity fraud). Thus, past studies may have underestimated the applied problem. In this thesis I examine face recognition performance for these challenging image scenarios and test whether the familiarity advantage extends to these situations. I found that face recognition was indeed even poorer for challenging images than previously found using cooperative images. Familiar viewers were still better than unfamiliar viewers, yet familiarity did not bring performance to ceiling level for challenging images as it had done in the cooperative tasks in the past. I investigated several ways of improving performance, including image manipulations, exploiting perceptual constancy, crowd analysis of identity judgments, and viewing by super-recognisers. This thesis provides interesting insights into theory regarding what it is that familiar viewers are learning when they are becoming familiar with a face. It also has important practical implications; both for improving performance in challenging situations and for understanding deliberate disguise.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgements

First I would like to express my sincere gratitude to my supervisor Dr Rob Jenkins, for his continued guidance and support, as well as for his enthusiasm for each and every project throughout my PhD.

I would also like to thank the members of the University of York FaceLab as well as my friends and family for all of their encouragement and helpful advice.

Finally, thank you to my disguise models – so much of this thesis would not have been possible without the phenomenal effort and dedication you put into creating your disguises. For that I am incredibly grateful. Your involvement helped make this PhD so enjoyable.

# Declaration

I declare that this thesis is my own work carried out under normal terms of supervision. This work has not been previously been presented for an award at this, or any other University. All quotations in this thesis have been distinguished by quotation marks and they have been attributed to the original source. All sources are acknowledged as References.

# Chapter 1 – General Introduction

## 1.1 Introduction: Why Face Recognition is Important

Face recognition refers to the ability to correctly verify the identity of a specific individual, often by comparing a 'target' face against other face images. Accurate facial identification is important because facial image comparison is the basis of many security infrastructures, such as passport control. Successful face recognition is often critical in the identification of criminal suspects, and avoiding miscarriage of justice.

**The Applied Problem**

The need for face recognition research became apparent following several high profile cases of mistaken identity. In England in the 1890s Adolf Beck was convicted of fraud and imprisoned as a result of erroneous face recognition. Beck was first imprisoned in 1896, after being repeatedly picked out in a police line up of face photographs as being the man responsible for defrauding over 20 women in different attacks. All women had reported that their attacker was a grey haired man who had a moustache. Beck's face was the only face in the line up to have both of these features. After serving his sentence Beck was released, but soon imprisoned after again being identified (based on eyewitness testimonies and line up scenarios) for more attacks similar in nature to those prior to his prison sentence. It was only when these similar attacks continued during Beck's second prison sentence that the real culprit 'Smith' whose real name was Meyer, was finally caught. Three of the five women who identified Beck as the culprit at the hearing prior to his second imprisonment were called in to view photographs of Meyer. All women admitted their mistake in their previous recognition of Beck and agreed that Meyer was the true attacker. Meyer had been following the original Beck case and had moved away to America until Beck served his sentence, returning to the UK and to his previous fraudulent activity following Beck's release but was unaware that Beck had been convicted a second time (Coats, 1999).

Similar high profile situations of mistaken identity still exist in the United Kingdom. On the 22nd July 2005, a case of mistaken identity cost Jean Charles de Menezes his life. The metropolitan police mistook Jean Charles de Menezes, a Brazilian electrician, for Hussain Osman who was a suspect involved in failed bomb attacks, which had been carried out in London the previous day. Several incidents led to the police shooting Jean Charles de Menezes eight times, believing that they were faced with Hussain Osman at the time of shooting. One of the contributors was erroneous face-matching - police had been given a photograph of Osman, and mistook Jean Charles de Menezes to be him (see Figure 1 for example images of the suspect and mistaken suspect) (BBC News, 2005; Cowan, 2005). This tragedy echoed the message that when face recognition goes wrong, very serious consequences can occur.

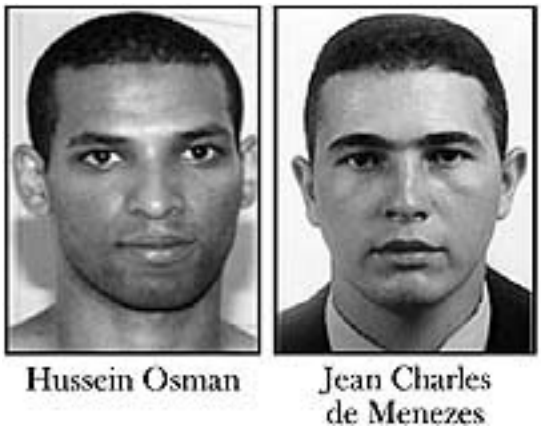


**Figure 1.1** Image showing the face of the real suspect, Hussain Oman on the left and the face of Jean Charles de Menezes on the right.

In addition to these famous cases, several studies highlight the weight that eyewitness face recognition can hold on a jury verdict. Wells et al. (1998) found that out of the first 40 cases where later DNA evidence excluded wrongfully convicted suspects as the culprit of a crime, 90% of these wrongful convictions had been strongly based on evidence from erroneous eyewitness testimonies. Additionally, Huff (1987) found that mistaken eyewitness identification of a convicted culprit occurred in 500 wrongful convictions that were later investigated.

Loftus & Doyle (1992) tested the weight given to eyewitness testimonies experimentally by creating a mock trial situation. Two separate sets of jurors took part, each set were presented with the same evidence, however one of the sets of jurors were also given evidence from an eyewitness testimony whereas the other set of jurors received no eyewitness testimony. When no eyewitness testimony was presented, 18% of jurors gave a guilty verdict. This jumped to a guilty verdict of 72% for those who received the eyewitness testimony. Further still, even when the eyewitness testimony was said to be inaccurate, 68% of jurors delivered a guilty verdict. Jurors tend to overestimate the accuracy of eyewitness testimonies (Brigham & Bothwell, 1983). It is important to understand whether these cases of mistaken identity are unfortunate anomalies in face recognition or whether they are indicative of a very serious problem – people may not be as good as they think they are at recognising faces. A great deal of research has been done on this, which will be discussed later in this chapter.

**Solutions to the Face Recognition Problem**

The examples and studies discussed above clearly demonstrate that there are situations where face recognition has been inaccurate in the past. Several procedures and technologies exist to aid identification. For example, passports, driving licences and university student cards feature an image of the holders face. This is intended to allow a viewer to make an identity judgment by comparing the image of the physically present cardholder with the image on the card that they are holding. If the images are considered a match, then the holder will be granted the access or purchase that they wish to acquire. If the images are not considered to match this would imply that the holder is using fraudulent identification, and the carrier would not be granted their access or purchase request.

Additionally, police line-up situations exist to try and reduce the number of erroneous face identifications in an investigation. In a line-up situation, witnesses of a crime view

the faces of several people, often of a similar appearance, one of whom is the suspect in a criminal case. The witness is tasked with picking out the suspect from the line-up array of faces. Line-up situation errors existed in several of the cases discussed above, demonstrating that this method of facial identification does not always render accurate results.

Closed circuit television (CCTV) cameras have become more prevalent in recent years meaning that more facial images are stored now than ever before. CCTV camera footage is used in many forensic investigations as the images can both capture the image of a suspect and track the suspect's movement. CCTV camera footage is however not always of good quality, and images available for comparison may come from multiple cameras, with breaks in the footage where CCTV is not in operation.

**Demands Experienced by Human Viewers**

Although there are several methods in place, which aim to ensure accurate identification of an individual occurs, each of the methods places demand on the human viewer who is required to make the identity judgment from the visual information available. Document holder to card image identity comparisons and CCTV camera footage evaluation require a viewer to make an identity judgment by *comparing* the images presented before them to determine whether the identities are the *same person* or *different people* across the multiple images. Line-up scenarios include a *memory* component – a witness has to compare the previously seen image of a face, which they have stored in memory, with the images seen before them. This memory component to the task may affect face recognition accuracy. Human ability to meet each of these demands is assessed below.

**Human Face Memory Ability**

The famous cases of mistaken identity discussed earlier involved aspects of erroneous face memory. To test face memory performance experimentally, Bruce (1982)

investigated human face memory performance for identical images, images changed in either pose or expression and images changed in both pose and expression. Participants in the study viewed images of 24 people, and were told to try and remember the face of the person in each of the images viewed as they would be asked to identify the images in a later recognition task, in which the identity may be shown in a changed image. Recognition accuracy was highest for the unchanged images (90% rate), followed by the images changed in pose or expression (76% hit rate) and finally images changed in both pose and expression (60.5% hit rate). It is important to note that in real world face recognition tasks the images for comparison are always changed between learning and viewing the face at a later stage. This study provides experimental evidence for poor face memory performance for changed face images. Alongside the example of famous cases of mistaken identity discussed above, it is clear that face recognition involving face memory is problematic.

**Perceptual Matching**

Initially, memory was thought to be the whole problem of face recognition. However, in recent years, identification by face-matching has become important. In face-matching the task is to compare two images—both of which are physically present – in order to make identity judgments. This is the scenario used in passport control, other identity checks involving photographic identification, and also by police to piece together an incident through comparison of multiple CCTV images. There is no memory component in this task as all of the information needed to make a decision is available in the images presented. However, it soon became clear that the problems of face recognition surpassed erroneous memory.

## 1.2 Familiarity & Face-Matching

Face-matching accuracy differs greatly for faces that a viewer is familiar with compared to performance for faces that are unfamiliar to the viewer (e.g Burton, Wilson, Cowan & Bruce, 1999). This section will examine face-matching performance for both unfamiliar

viewers and familiar viewers in turn and then explore the idea that familiarity is more than a binary variable; investigating research which treats familiarity as a graded concept (Clutterbuck & Johnston, 2002,2004,2005).

**Unfamiliar Face-Matching**

People are extremely accurate at matching two identical images of a face (Bruce, 1982). Face recognition, as required by the security and forensic situations outlined above, relies on the ability to match or identify a mismatch between different images of faces. The literature covers several unfamiliar face-matching scenarios including person to photograph matching (Kemp, Towell & Pike, 1997; White, Kemp, Jenkins, Matheson & Burton, 2014), police line up scenarios (Burton, Wilson, Cowan, & Bruce, 1999) and paired image comparisons (Megreya & Burton, 2006, 2007; Burton, 2010).

**Unfamiliar Live Face-Matching**

The comparison of a face photograph to a physically present face is the method of identity verification that matches that of passport control, building access, shop sales and many other security situations. There is no memory component in such a task as both the physically present cardholder and the identification card are available for direct comparison. Kemp et al. (1997) tested performance accuracy for matching an ID card holder (physically present) to a photo on the ID card. The experiment took part in a large supermarket in the outskirts of London where participants in the study were experienced supermarket cashiers. The cashiers' task was to determine whether the photograph on the identification cards portrayed the holders. Each cashier judged physical appearance against the identity cards for an average of 44 shoppers with whom they interacted at the checkout (interaction was as in a normal shopping scenario). Identification performance was poor - the cashiers performed with 67% accuracy. More than half of fraudulent cards were accepted as the holder's true identity. In this experiment identification card photographs were taken just six weeks before the experiment. In the UK, passport photographs are valid for up to ten years before the photograph needs updating. This

means that the cashier task could be even easier than real life situations of cardholder to image identity matching as a face may look even more different over longer periods of time.

While performance was poor for the supermarket cashiers, one might expect trained personnel who work in security to excel at a live person to image matching task. White, Kemp, Jenkins, Matheson & Burton (2014) recently tested passport officers face recognition ability in a task similar to Kemp and colleagues (1997) study. Passport officers were tested on their ability to match live subjects to a passport style photograph that was either of the same identity or showed an image of a different identity. Just 16 female, and 16 male student volunteers took part as the photo holders and identities for this experiment. Different person trials were created though pairing together the most similar images from the images provided by the volunteers. As the volunteer pool was small, this greatly limited the ability to provide an extremely convincing foil. Additionally, all photographs were taken just two weeks before the experiment was carried out. These conditions may make the experiment easier than a real life situation would be, as in real situations of identity fraud people may have a large sample of available different person passports to chose from, and additionally for same person trials, people may look more different in their real passport photograph than they did in the image used in this task as less time had passed since the time of photograph, leaving less time for natural changes as a result of aging. Nevertheless, performance in the matching task was poor. On average, passport officers made errors in 10% of cases, with 14% of different identity pairs being mistakenly accepted as same people. Years of experience and training in matching faces made no difference to performance. Despite overall levels of poor performance, there were some individuals who performed very well on the task (see Figure 1.2). A photo-to-photo matching test was then conducted to compare passport officer performance with that of undergraduate student participants. Student's performance was indistinguishable from that of the passport officers. This study highlights the difficulties of live-person to image matching, and also demonstrates that passport officers, a highly trained group, are no better than undergraduates at face-matching.

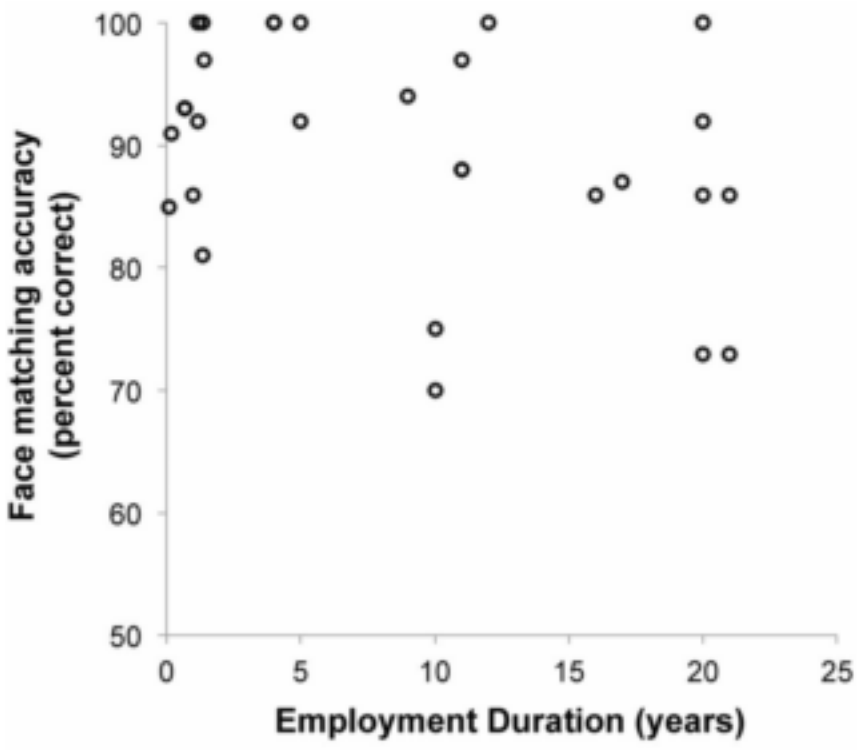


**Figure 1.2** Face-matching performance of passport officers in the study by White et al. (2014). Some officers perform with very high performance regardless of their employment duration.

**Line-ups**

In addition to live person to image matching, poor face-matching performance has also been reported in experiments that replicate the police line up scenario. Bruce, Henderson, Greenwood & Hancock (1999) created a face-matching task based on a police line up situation. Participants had to match a given photograph of a face to one of the face images from a line up of 10 face images presented in an array below the target face (see Figure 1.3). Participants were told that the target face may or may not be present in the array line-up, as would be the case in a real criminal line-up situation. Notably, unlike a real line-up situation, no memory component was involved for the viewer of the faces. All face photographs were taken on the same day, in the same lighting, and in the same pose. The only deliberate modification between images was that the target photograph was taken with a different camera to the array photographs. Despite the consistencies between the target photograph and the correct match in the array line-up, performance

was poor. For both the target-present and target-absent array scenarios, accuracy was 70%. Performance accuracy for correctly matching a target to the same identity present in the array remained poor even when participants knew that the target face was present in the array (Bruce et al. 1999; Bruce, Henderson, Newman & Burton, 2001).

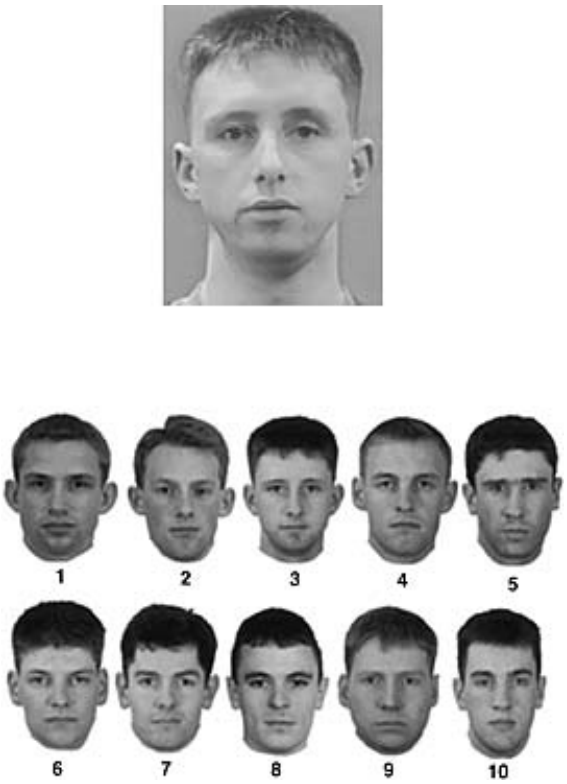


**Figure 1.3** Example of line stimuli taken from Bruce et al. (1999) line up task. The correct match of the target face is face number 3.

### Paired Face-Matching

More recently, work on face-matching ability has focused on paired matching paradigms, in which viewers make identity judgments (same or different) for two faces presented simultaneously. The viewer's task is to say whether the two faces in the pair presented depict the same person or different people. Face-matching by image is a good model of real world tasks and has important practical applications. For example, police investigations often require accurate matching of an image taken from CCTV footage to another image of the suspect, such as a mug shot. Alternatively, two different face images taken from different CCTV footage may be compared in order to track a person's

movement. As in the experiments described above, the face-matching paradigm involves no memory component – both images are viewed simultaneously and the viewer's task is to say whether the faces presented are of the same person's face or are faces of two different people.

To provide a standardised measure of face-matching ability, Burton, White & McNeil (2010) designed a task known as the Glasgow Face-Matching Test (GFMT). The images used to construct this test were taken at a constant viewing distance and same camera angle, under constant lighting conditions, and the expression of the face was unchanged across photographs of the same identity (see Figure 1.4 for examples). Photographs of the same individual were taken only minutes apart, with no deliberate attempt to change the model's appearance in any way between photographs. Thus, images should have been of optimal condition to allow accurate comparison. Different person trials were created by pairing the faces with the most similar other person's face image out of those available. Two versions of the GFMT were created – the full version (160 items), and a short version consisting of the 40 most difficult items from the full version. Participants made over 10% of errors in the full version. This performance is considered low, especially considering the controls taken to make the images as similar to each other as possible and given that participants experienced no time limit while completing the task. The short version of the task rendered poorer results, with participants making nearly 20% errors.

**Figure 1.4** Examples of face pairs taken from the GFMT (Burton, et al. 2010) Top row show different identity pairs, bottom row are same identity pairs.

In summary, these studies successfully and repeatedly demonstrated the real world problem of poor face-matching in a lab environment. However, these findings go against most people's intuition. People tend to believe that they are good at face recognition (Jenkins & Burton, 2011). It is important to remember the vast majority of face-matching that people are faced with on a daily basis involves familiar faces. People experience regular and repeated success in recognising family members and friends, and do so with very little difficulty.

**Familiar Face Matching**

People are extremely good at matching familiar faces, even when image quality is poor. To test this, Burton, Wilson, Cowan & Bruce (1999) designed a face-matching task using poor quality CCTV footage (see Figure 1.5). This matching task was presented to people who were familiar with the faces in the CCTV footage, and also to a group of unfamiliar viewers. Familiar viewers could match the faces in the task with almost perfect levels of accuracy, whereas unfamiliar viewers performed at chance level or highly inaccurately on the exact same task (Bruce, Henderson, Newman & Burton, (2001); Burton, Wilson, Cowan & Bruce (1999).

**Figure 1.5** Examples of face pairs in the matching task involving poor quality CCTV footage in the experiments conducted by Burton et al. (1999). Familiar viewers performed with high accuracy when matching the face pairs.

This familiarity advantage can be achieved quickly. Megreya and Burton (2006) familiarised participants with faces by showing participants 30 second long video clips of the 40 identities that the participants were told to try and learn as they would see the faces again in a subsequent face-matching task. Even after this brief familiarisation with a target face, participants performed better in the task in trials involving familiarised faces over novel faces. However this advantage only existed for upright faces. When the familiarised face images were inverted, participants performed better on the unfamiliar upright faces than the inverted versions of the familiarised face images. This is likely due to the range of experience that a viewer has with a face and is a topic which will be addressed later in this chapter. Megreya & Burton (2006) concluded that unfamiliar faces are "processed for identity in a qualitatively different way than are familiar faces". The authors suggested that unfamiliar faces are processed in a similar way to other objects whereas we have a special ability for faces when they become familiar, which relies on a different processing system.

**Familiarity as a Graded Concept**

Many past studies have classified faces as either familiar or unfamiliar. Recently, experiments have demonstrated a graded nature to familiarity, with familiarity with a face increasing as exposure to that face increases. To test for a graded nature of familiarity, Clutterbuck & Johnston (2002) divided the familiarity of faces viewed in their study into 3 levels, rather than treating familiarity as a binary variable. Faces in Clutterbuck & Johnston's (2002) study were described as being highly familiar, of medium familiarity and unfamiliar. Familiarity distinctions for these celebrity face images were made by a group of eight independent raters. The high and medium familiarity group were created based on the mean familiarity ratings for each face. Unfamiliar faces were faces that were not famous and were not known by the participants. The results demonstrated that participants were fastest at accurately matching full faces to internal features of a face for the faces for the highly familiar faces in the study (Figure 1.6). Speed of correct response decreased as familiarity level decreased across each of the three familiarity groups. All of the faces used for the familiar and medium familiarity groups belonged to celebrities. By breaking familiarity into 3 levels this study begins to separate the level of familiarity of these celebrities which begins to address the issue that not all celebrities are equally well known. However, this method fails to address that different participants in the study may be more or less familiar with different faces to each other.

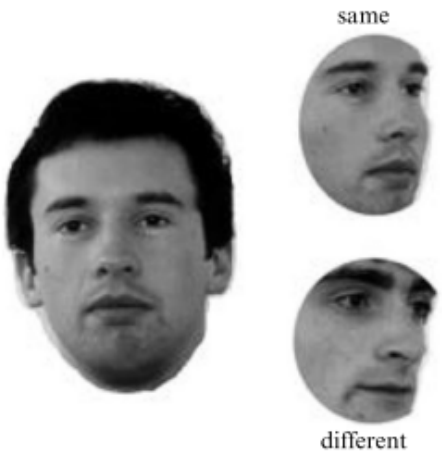


**Figure 1.6** Examples of full-face image on the left was to be matched with either of the internal face images on the right. Familiarity aided matching accuracy on this task.

Clutterbuck & Johnston (2004) also demonstrate a graded familiarity advantage for accuracy in a gender judgment task. The authors again distinguished between three levels of familiarity but constructed the familiarity groups in a different way to their previous (Clutterbuck & Johnston, 2002) study. Faces in the gender judgment tasks were either previously familiar (know to the participant before the experiment), of acquired familiarity, or unfamiliar (never seen before the testing phase of the experiment). Participants achieved acquired familiarity by viewing sets of previously unfamiliar images for two second periods, with each face image viewed a total of ten times. The results of the study were that participants were best at the identity judgment task for faces which they were previously familiar with (familiarity was measured by eight independent raters as in Clutterbuck & Johnston, 2002), and gender judgement performance was of higher accuracy for acquired familiarity faces than for the unfamiliar faces. This acquired familiarity method gave a similar pattern of results for face-matching speed (Clutterbuck & Johnston, 2005). Participants were fastest at correctly rejecting mismatch faces when they were previously familiar with them, followed by those for which they had learnt as part of the experiment, and slowest for faces which were unfamiliar at test. The three experiments conducted by Clutterbuck & Johnston (2002, 2004, 2005) suggest that face recognition is influenced not only by whether a face is familiar or not, but by the level of the viewer's familiarity with the face.

## 1.3 What Information is used to Recognise a Face?

The literature has shown that people get better at recognising faces when they learn the faces in question, but it is unclear what information from an image viewers actually use to make identity judgments of faces.

There has been much debate in the face recognition literature over which information from a face allows the face to be recognised. The main lines of debate over the information used within a face to aid identification, focus on a *featural* versus *configural* processing account. *Featural* processing involves scanning the face for features (e.g. eyes,

nose, mouth), and then using the information from the detail of these features to decide whether the face is known. *Configural* processing on the other hand, argues that the spatial layout between features is what is important for successful face recognition. According to the holistic processing theory, faces are identified 'as a whole entity', with both a match in configural and featural information being important for recognition (Sergent, 1984; Bruce, Hellawell & Hay, 1987). Studies have generally taken the approach that if a manipulation applied to a face image slows down and/or impairs face recognition, then whatever was changed about the face was important for the face recognition process (Donders, 1868/1969).

**Featural Accounts**

Bradshaw & Wallace (1971) provided one of the first studies to suggest that face recognition is based primarily on the recognition of facial features. Their experiment comprised of images of identikit faces - faces made up of face elements [individual features], which can be combined to make a face. This method is used by police to create an image of a suspect based on a witnesses' description of the suspect's face. Participants in Bradshaw and Wallace's (1971) study viewed pairs of identikit faces and judged whether the image pairs presented showed the same face or different faces. Different parings differed in the number of shared features in the two images. For different person trials, accurate identity judgments were made more quickly as the number of differences in features between the faces increased. On the basis of this result it seemed that participants scanned the face images until they found a mismatch of features across the image pairs. Thus, it was concluded that the results were best explained through a featural account of face recognition. An earlier experiment by Smith & Neilsen (1970) supported this result. They reported that as the number of featural differences between pairs of faces was increased, there was a decrease in the time needed to tell the two faces apart. This suggests that participants were using a feature comparison technique to aid their identity judgment of the face images. These studies looked only at performance for unfamiliar faces, therefore they do not provide information on what is learnt during the familiarisation process with a face.

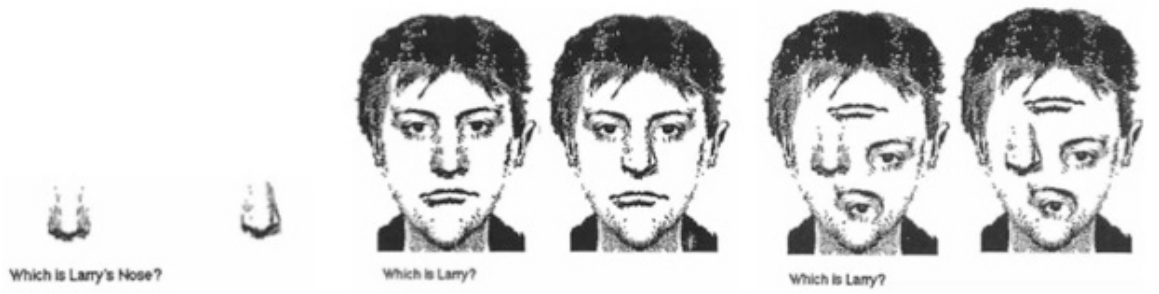


**Figure 1.7** Example of test stimuli used by Tanaka & Farah (1993), which shows examples of isolated features, intact faces and scrambled faces.

Further evidence for a featural processing account comes from studies involving scrambled faces. Scrambled face images contain all the featural information as in an intact face image, however the features appear in a jumbled order. Tanaka & Farah (1993) tested whether participants were able to recognise learned identities from scrambled face images, or even from viewing a feature from a face in isolation of the whole face. They reported that participants made identity judgments with moderate accuracy in the context of a scrambled face and also when features belonging to a previously seen identity were viewed in isolation of the rest of the face (see Figure 1.7). This suggests that features, regardless of their configuration or even the presence of the rest of the face, provide information which aids face recognition (Bruyer & Coget, 1987; Tanaka & Farah, 1993).

**Configural Processing**

In addition to the research that argues for the importance of featural information in face recognition, there is a wealth of research that configural processing is important. Theories of configural processing suggest that the layout of features within a face (1$^{st}$ order configural processing), specifically the metric distances between features in a face (2$^{nd}$ order configural processing) are learnt and used in face recognition. For example, Haig (1984) tested the effect of digitally editing the distance between features of a face, without editing the features themselves in any way, on face recognition performance. Haig (1984) found that even when featural information was present in a highly accessible

form (see Figure 1.8 for example stimuli), recognition for unfamiliar faces was severely dampened when subtle changes were applied to the relationship of the layout of the facial features. This supports theories of configural processing.


**Figure 1.8** Example images of one identity with features changed in configuration across the different images taken from the study by Haig (1984). Features themselves remain intact, although the exact distances between features are changes.

Many studies on this topic are based on a similar logic; any manipulation that is seen to reduce recognition accuracy or speed is thought to be important for the recognition process. It is believed that configural information is harder to access in an upside down image than in the upright form that people have everyday experience with (Yin, 1969). Familiar orientation had been found to be important for past studies involving object and letter recognition (Henle, 1942; Ghent, 1960), and it is argued that the familiarity factor for upright orientation also exists for face recognition (Yin, 1969). Therefore, if configural information is important for face recognition, performance accuracy would drop in an inverted face memory task, as configural information is more accessible in an inverted than upright face image. Yin (1969) found exactly that result. In general, people show a bias for remembering upright objects, with various upside-down objects proving more difficult to remember (Goldstein, 1965; Hochberg & Galper, 1967), however, faces were

disproportionally affected by inversion, with poorer results for remembering faces than objects in the inverted image memory task (Maurer, Grand & Mondloch, 2002; Rossion, 2009). Poorer recognition performance for inverted than upright faces suggests a role for configural information in face recognition. Again, this finding was for unfamiliar faces.

Despite the arguments for configural processing provided by the studies discussed, the usefulness of the content has been queried. Several studies show that configural information in a face can be changed, and faces still successfully recognised. Hole, George, Eaves & Rasek (2002) investigated whether familiar face recognition was affected by drastic changes in the featural layout of the face. They reported that faces could undergo vertical stretching of twice their original height, with no change in face recognition performance for these face images (see Figure 1.9 for example of image stretching). Bindemann, Burton, Leuthold & Schweinberger (2008) added to this finding, as they demonstrated that the N250r ERP response to faces is unaffected by stretching of the face. These findings suggests that neural processes involved in face recognition, and also behavioural face recognition, are unaffected by the configural changes caused by stretching, suggesting that consistent configural information is not important for these processes.

**Figure 1.9** Example stimuli from Hole et al. (2002) showing original image, vertical stretch of 150% and 200%.

The configural account has also rendered failure in practical application. Configural processing has failed to provide adequate facial identification results in both early

automated face recognition systems (Kelly, 1970; Kanade, 1973), and more recently Kleinberg, Vanezis, & Burton, (2007) presented evidence that anthropometry methods fail to identify targets in forensic investigations.

**Are Both Important?**

Collishaw & Hole (2000) investigated the effect of a range of face image manipulations on face recognition performance for both familiar and unfamiliar faces (Figure 1.10). Two of these manipulations – scrambled and inverted (upside down) faces – disrupted the configural information in an image, but kept the features themselves intact. Another – blurring – made it very difficult to access the featural information in a face. It was found that the identities in the image, regardless of the viewers familiarity with the identity, could be recognised 'much of the time' in all of these situations. Thus, suggesting that as long as one of these processing methods is available, face recognition can take place. Further image conditions – blurred and scrambled, blurred and inverted – were recognised at levels of around chance. This finding reaffirms the need for one of these routes of recognition to be present for successful recognition to occur.

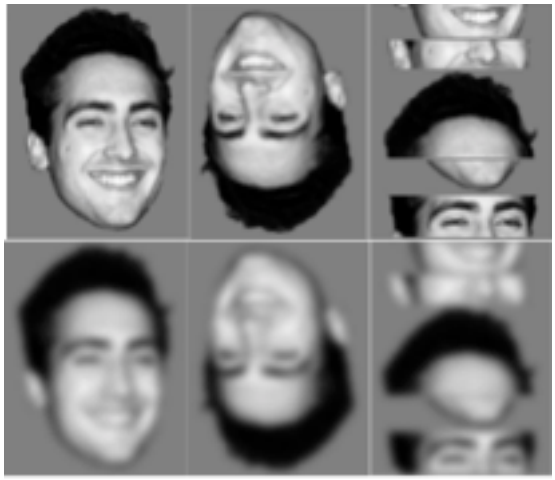

**Figure 1.10** Example stimuli from Collishaw & Hole (2000). Top row, from left to right images show the following image conditions: intact, inverted, scrambled. Bottom row images show: blurred, blurred and scrambled, blurred and inverted.

In summary, evidence points to neither featural or configural accounts providing a full answer to the face recognition problem for both unfamiliar and familiar face recognition. Images can be difficult to recognise when featural information is present but configuration is changed, and recognition performance can also be poor when configural information is intact, such as in photo-negatives and sketch drawings (Galper, 1970; Davies, Ellis & Shepherd, 1978; Rhodes, Brennan & Carey, 1987; Bruce, 1992; Kemp, 1996). It is also possible that recognition relies somewhat on texture information – something ignored by the configural and featural accounts (Kemp, 1996; Bruce, 1992).

## 1.4 Learning Variability

It is possible that what is being learnt when becoming familiar with a face, is not in fact a specific factor such as the exact details of features themselves or of facial configuration, but instead an accepted range of appearances for any given face. People are very good at recognising face images for people they are familiar with, providing any new image is similar to those previously experienced, but are poor at recognising people in situations beyond this experience (Bruce, 1994). People vary in appearance naturally across multiple images. This can be a result of changes in expression, pose, hairstyle, skin complexion, clothing and many other variables. Bruce (1982) demonstrated changes in pose or expression between items learned and presented at a later recognition test, lowered recognition accuracy for unfamiliar viewers. Unfamiliar viewers could rely only on the information from their one previous exposure with the face image, and could not generalise this exposure to changes of expression and viewpoint. Familiar viewers performed accurately on the task despite these image changes, as they could call on a large number of previous encounters of the face under different conditions to aid their identity judgments (Bruce, 1982; Burton et al. 1999). Familiar viewers cannot however generalise past experience with a face to aid recognition in the case of novel face transformations. For example, familiar viewers have been found to perform poorly in recognition tasks of photo negative faces and images changed in pigmentation (Bruce & Langton, 1994) – these changes fell outwith the previous range of experience acquired for the familiar faces.

Further evidence to suggest that familiar viewers store an accepted range of experiences from a face come from studies on facial configuration. Bruce, Doyle, Dench & Burton (1991) presented participants with a series of face images which were generated using facial image software Mac-a-Mug. Different versions of the same face image were created to provide various configurations of the same face (e.g. the features could be moved up, down, further apart or closer together). Participants were later presented with a series of test image pairs, and asked to identify which image from each of the pairs was identical to an image that they had seen earlier. Accuracy was at ceiling levels for recognition of an identical face. Faces where configurations lay outwith that in the previously seen face images were not recognised at test. However, the central image of those previously viewed (although never itself presented), was also rated as highly familiar, demonstrating that familiar viewers would accept face images which fell inside the range of face images they had experienced for a face, but reject any faces which fell outside this experienced range (Bruce et al., 1991). More recently, Sandford and Burton (2014) asked participants to resize an altered image of a face to make the face 'look right'. Familiar viewers were no more accurate than non-familiar viewers at this task. If configuration does not remain constant for faces across various images, then it makes sense that there would not be just one accurate configuration, but a range of accepted configurations that could be stored.

The effect that learning variability has on face-matching accuracy is demonstrated in a study by Jenkins White, Montfort & Burton (2011). First they highlighted the difficulty that variability between photographs of the same face causes in unfamiliar face-matching. Jenkins and colleagues (2011) challenged participants to separate images of cards into piles according to identity. Images of the same identity were to be placed in the same pile. There were in fact only 2 identities - 20 images of each. The 2 identities were Dutch celebrities who were well known in the Netherlands but unfamiliar faces for British viewers. Unfamiliar (British) participants averaged rating of the number of identities present among the cards was 7.5. Familiar viewers (Dutch participants) were then asked to do the exact same task, and easily and accurately separated the faces into the two identities present. This study highlights the range of variation that familiar viewers have

stored for an identity and the advantage this holds in identity matching. Figure (1.11) demonstrates some of the many ways that the same faces can vary across images.

**Figure 1.11** Example of the card-sorting scenario. Figure shows 20 images of two different identities. Familiar viewers find this task easy whereas unfamiliar viewers generally believe that there are more than two identities present.

### Face Space Theory

One influential concept for understanding face recognition by variation is the theory of face space. It is possible that when a new face is encountered, or a previously viewed face takes on an appearance different to the way it looked before, the mental representation for that face image is then stored in that person's '*face space'*. Valentine (1991) was first to propose the face space model for faces. Valentine argued for a single, although multidimensional, face space, in which faces for all identities are stored. Faces could be stored as a single averaged image for each identity, or individual face images could all be stored within this face space. Valentine believed that faces were organised within this space by dimensions specific to faces. Specifically, faces were proposed to cluster within face space in ways which explain phenomena of face recognition including the own race bias, and typicality effects (Valentine, 1991). There are however shortcomings of this proposal, for example Burton & Vokey (1998) argue that most of the faces within

Valentine's proposed face space are located away from the centre of face space, therefore there are very few 'typical' faces.

An alternative formulation of face space would be that instead of one face space encapsulating all faces, in order to account for variation within each individual face, each face may have its own face space. Each face space could be thought of as a multidimensional 'bubble' that encapsulates all the ways and range of ways that *one* face is accepted to look. Any new image would then be compared against the representation held in face space, for that face, and against all other stored identities' face spaces, in order to make an identity judgment. Expertise would be acquired one face at a time, not for all faces as a class of stimuli.

This formulation is in line with the familiarity advantage, as familiar viewers have had more encounters with a face or been exposed to the face across more varied viewing conditions than an unfamiliar viewer. Thus, when a person is becoming familiar with a face they are learning the different appearances that the specific face can take, and this refines the face space held for that individual. If, as I suggest, identity judgments are made by comparing a new image with the representation of a face stored in that person's face space, then identification accuracy should increase in alignment with refinement of face space, i.e. familiar viewers will be more accurate as they have a more refined face space than unfamiliar viewers. This means that false match (*different identity*) items are more likely to fall *outwith* the range of accepted faces for the target face for familiar than unfamiliar viewers, and true (*same identity*) images of target will be more likely to be accepted to fall *within* the accepted range for the target face. It appears that exposure to greater variation with a face increases the likelihood of later correct identity judgments being made (Menon, White & Kemp, 2015).

Within person variability is a strong component of this face space model, as an identity specific face space would encompass the idea that people's appearance naturally changes

across images. Sometimes people can look unlike themselves, or similar to another person, just by chance. This *incidental* variation between faces is very different to deliberate change in appearance. Whilst people can look incidentally unlike themselves, or like another person they can also make a *deliberate* attempt to look unlike themselves (evasion disguise) or like another person (impersonation disguise). These appearance changes, both *incidental* and *deliberate* (disguise) will be a key theme of challenging images explored in this thesis.

## 1.5 Face Recognition in Challenging Situations

Past research has demonstrated that people are very good at matching two identical images of a face and that problems instead arise when dealing with matching multiple images of the same face (Bruce, 1982). When the target faces are unfamiliar to viewers, performance on such tasks is very poor (Kemp et al. 1997; Burton et al. 1999; Megreya & Burton, 2006,2007; White et al. 2014). However familiarity with the target faces aids face recognition accuracy in a range of situations (Burton et al. 1999, Bruce et al. 2001, Megreya & Burton, 2006, Jenkins & Burton, 2011). The studies discussed up until this point have tested face recognition where the people who are photographed to provide images for identity comparisons *cooperate* with the identity effort, i.e. for same person trials the models made no deliberate attempt to change appearance across the different images taken. Furthermore, different person trials were created by selecting the most similar face images from a small number of available images, therefore the different identity images may not look all that similar to each other. Face-matching performance could be even worse for more challenging identity matching scenarios.

As demonstrated above, people can look very different to themselves across different images (Jenkins, 2011). There are also situations where people naturally look very similar in appearance to another person. In addition to incidental change, in real world scenarios people may have strong incentives to create deliberate changes to their appearance - either to evade their own identity (*evasion*) or to impersonate someone else (*impersonation*). There may be reasons for a person to hide their identity (evasion) for example if they had been banned from a place but wanted to gain access. Impersonation

disguise is also a serious issue. Criminal activities including illegal immigration, the smuggling of drugs, weapons or stolen goods, human trafficking and terrorist activity may involve the use of stolen identity documentation in order to cross borders as someone else. People might choose a passport based on natural similarities with the face. It is also possible that the new document holder will make an effort to look like the photographed face on the document in order to reduce suspicion and successfully pass borders using the stolen identity. Such attempts are made quite frequently, for example, in 2010, 359 people were found guilty in the UK for possessing false or improperly obtained ID. Under the 2006 Fraud Act an additional 7,319 were charged for dishonestly making a false representation to make gain for oneself or another (Home Office, 2012). There have been some attempts to investigate the effect that i) natural/incidental image manipulations; and ii) deliberate disguise manipulations, have on face recognition performance. These will be discussed below.

**Natural Image Manipulations**

It is likely that there are certain image manipulations that increase the likelihood of these challenging image situations arising. Manipulations including changes in pose and expression have already been found to make identity judgments more difficult across images (Bruce et al. 1982).

It is possible that another simple image manipulation, changing camera-to-subject distance between comparison images, could make same person identity judgments more difficult than comparing images taken at the same distance. Harper & Latto (2007) showed that changing camera-to-subject distance changed perceived weight judgments of a face. This manipulation has not yet been experimentally explored with reference to face recognition ability, however the results suggest that images changed in camera-to-subject distance would be more difficult to 'tell together' as perceptions of the face changed as a result of camera-to-subject distance change.

**Deliberate Disguise**

Most of the research on recognising disguised faces has focused on changes to featural information in the face, with reference to evasion disguise. Patterson & Baddeley (1977) investigated whether face recognition was affected by disguise presence. In this study participants viewed (with the intent of learning) a series of face images. Participants were later shown a series of test images, some of which showed the previously seen identities in disguise, and were tasked with recognising previously seen identities from these images. Disguise manipulations reduced recognition accuracy to approximately chance. This task was an image memory task, where the exact images were used at test and learning, except for those images in the disguise condition. Disguised images included the addition of props to a face, such as wigs, glasses, fake moustaches and beards.

Terry (1993) found that a more specific and simple form of disguise, the addition of glasses to a face, was in itself detrimental to face recognition performance. However, this study reports an effect of eyeglasses on face image memorability, rather than an experimental disguise manipulation. Participants viewed images of faces (some with glasses and some without) and were later asked to identify the previously seen images amongst distractor images. There was no addition or removal of glasses between images. This was an image memory task rather than a face recognition experiment. In a later experiment Terry (1994) approached the eyeglasses manipulation from a more controlled angle. This time, participants learnt face images and were then tested on face images that could have had glasses or beards added or removed to the image. The removal of glasses on a person who had initially been presented wearing glasses, and the addition of a beard lowered recognition accuracy (Terry, 1994).

Furthermore, Righi, Peissig & Tarr (2002) suggested that some disguise manipulations were more detrimental to memory performance than others. They reported that recognition performance was hindered if the face image changed in any way between the learning and test phase. However, recognition was significantly worse when the disguise

manipulation involved a change of hairstyle or removal of glasses, compared to when just glasses were added to a face.

The effectiveness of different types of disguise may depend on the viewer's familiarity with the disguised face. External features of a face have been found to be of particular importance in unfamiliar face recognition (Bruce et al. 1999; Bonner, Burton & Bruce, 2003; Megreya & Bindemann, 2009), whereas internal features may be of greater use in familiar face recognition (Ellis, Shepherd & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985). To date, disguise studies have not looked at the internal/external feature manipulation in terms of effectiveness of disguise type and familiarity with the disguised face.

The studies of disguise discussed so far have all involved a face memory component. Dhamecha, Singh, Vatsa & Kumar (2014) provide the only published study to date to investigate people's ability to face match disguised faces. Photographic models in the study were given a range of accessories which they could use as they wished to disguise themselves. Participants then completed a printed questionnaire, which showed pairs of face images, and were tasked with deciding whether the images in the pair were of the same person or not. Participants viewed faces of both the same and different ethnicity to themselves, and also familiar and unfamiliar faces. Highest accuracy rates were found for same ethnicity, familiar faces. This study provides a very interesting first look at face-matching ability for disguised faces, in particular it is interesting that familiarity aided recognition performance. Although this study provides a recent focus on the problem of face-matching with disguised face, the stimuli in the study may not be realistic (the sorts of manipulations that a person would naturally chose to disguise themselves) and certainly not undetectable disguises (see Figure 1.12). Many of the disguise manipulations occluded features of the face. Occlusion disguise would not be effective in evading identity in all situations, for example passport security checks often request that items of occlusion are removed from a face during an identity check. Furthermore, this study focuses exclusively on evasion disguise. This is a very interesting question to address as

people may have very strong reason to not be identified as their true self, however evasion disguise only covers half of the disguise problem. If a person is travelling on a stolen passport, it is more likely that they would attempt to make themselves look specifically like the person on the passport rather than only trying to hide their own identity. Given the threat of successful disguise to security it would be useful for studies to investigate realistic disguise for both evasion and impersonation situations.


**Figure 1.12** Example stimuli in the Dhamecha et al. (2014) face-matching task.

In summary, where disguised face recognition performance has been investigated, focus has been exclusively on evasion disguise (when a person changes their own appearance to look unlike themselves). Stimuli used in these studies have focused on simple disguise manipulations, mostly the addition of props to occlude parts of the face. Furthermore, experiments have compared disguise face recognition performance with exact image matching which does not address the question of whether disguise impairs performance compared to performance for recognising across different undisguised images of a face.

## 1.6 Overview of Current Work

There has been a great deal of previous work on human face-matching performance, however these studies have generally looked at performance in tasks where the models used to create same identity stimuli have cooperated with the identity effort and different person image pairs have been constructed from a small number of available

images. Performance could be even worse for more challenging stimuli. There has been some past work on the challenging case of deliberately disguised faces, but this has been limited to investigating on the case of evasion disguise, generally through purely the addition of props to a face.

This thesis will investigate face-matching performance for challenging images, including a more thorough examination of deliberate disguise. Past research has established that people perform poorly when matching unfamiliar faces, and familiar viewers perform very well even when images are of degraded quality. These findings come from experiments that have tested performance accuracy for matching cooperative stimuli. In reality, there are situations when people naturally look a lot like somebody else, and also situations where people look naturally different across multiple images of themselves. The images that result from these instances of natural similarities between identities or natural differences in appearance of own identity, would create a far more challenging task than the cooperative stimuli` tasks created to date. Further to these naturally occurring instances, people can make deliberate attempts to evade their own identity or to impersonate another identity. The effect of both incidental and deliberate changes in own appearance, and incidental and deliberate similarities with someone else's appearance will be tested in this thesis. Ways to improve performance will also be investigated with the effect of familiarity on face matching performance remaining a key theme throughout. Image manipulations and performance on face-matching tasks for challenging stimuli will help to address theoretical questions regarding the methods used for identity judgments.

The second chapter of this thesis explores whether face-matching accuracy is even poorer for challenging stimuli (ambient same identity images, and very similar different person image trials) than the poor performance found previously for matching tasks based on cooperative stimuli. To increase task difficulty and match the challenging image conditions often encountered in forensic investigations, images are degraded in quality

through image pixelation across a series of three experiments. The effect of familiarity on task performance is explored and treated as a graded rather than binary variable.

Chapter three acknowledges the problem of reducing image quality on face-matching accuracy that is established in Chapter two, and investigates ways of improving performance. Three techniques are tested, i) image manipulations; ii) data analysis; and iii) observer effects. These techniques are first investigated in terms of any improvement gain they bring when used alone. The effect of combining methods is also examined.

Chapter four investigated whether camera-to-subject distance changes influence face recognition, with an aim of investigating theory behind facial recognition, whether face-matching performance is impaired, and finally whether perceptual constancy methods exist to facilitate identification across distances. The chapter tests the effect of changing camera-to-subject distance on the facial configuration of a face as measured from an image, and also whether any changes result in difficulties in identity judgments. This is an incidental appearance manipulation, and may make the same identity look different across multiple images, and perhaps even different identities appear even more different. Performance accuracy for both familiar and unfamiliar viewers is again tested and perceptual constancy to help deal with distance induced changes to a face is explored.

The final chapters directly address deliberate changes in appearance through the creation of a new disguise face database, which encapsulates both evasion and impersonation disguises and also non-disguised images of these faces. Chapter five tested the effect of disguise on face matching performance and established effects of disguise type. The effect of familiarity was investigated and also whether unfamiliar viewers performance accuracy could be improved if they were informed that disguises might be present. Chapter six is an exploratory chapter, which aims at further understanding what people do to create evasion and impersonation disguises, and which of these approaches are

effective. This chapter concludes with an experiment that tests for differences in perceived personality judgments across non-disguised and disguised faces.

Taken together, performance accuracy for the challenging stimuli investigated is even worse than performance in previous face matching tasks that were constructed from cooperative stimuli. There are important distinctions between natural and deliberate efforts to not look like oneself and to look like another person. This is particularly evident in the study of deliberate disguise. Evasion and impersonation disguise cause different levels of face matching difficulty and the disguises themselves are achieved through the use of different methods. Familiarity aids performance in all tested challenging face-matching tasks, and there are several other methods that can be used to improve performance which do not rely on familiarity with the faces concerned.

# Chapter 2 – Familiarity & Challenging Faces

## 2.1 Chapter Summary

This chapter looks at the role of familiarity in a naturally difficult identification task – specifically matching similar faces. In Experiment 1 a graded familiarity advantage is reported, with participants being poor at the matching task for unfamiliar faces and much better for familiar faces. This graded familiarity advantage survived even when the images were pixelated to eliminate fine scale information in the images (Experiment 2) but began to break down under coarse pixelation (Experiment 3).

Pixelation makes featural and configural information difficult to access. The observed advantage of extremely familiar faces even under coarse pixelation suggests that other information besides fine scale information in the images was being used to support the required discriminations.

## 2.2 Introduction

Facial identification is a task people often assume they are good at (Bruce, 1988; Jenkins & Burton, 2011). This belief is likely held because our everyday experiences of face recognition with familiar faces cause us little difficulty. People can very easily identify family members, friends, and celebrities across a wide range of image conditions including different angles, different lighting, changes in pose and even over images taken years apart (Bruce, 1982; 1994; Jenkins, White, Van Montford & Burton, 2011). Confidence in human face recognition ability is so high that facial identification forms the basis of many major security systems worldwide. For example, passport control verifies personal identity by comparing photographic identification documents against the face of the holder (physically present).

Experimental evidence suggests that this confidence is misplaced. Decades of research has shown that people make frequent errors in face identification tasks when the tasks involve unfamiliar faces. This is problematic as the security situations relying on face recognition as an identity verification method generally involve unfamiliar faces, not familiar faces. Experiments have tested human face-matching performance in a variety of ways, including paired face-matching tasks, line up arrays and live person to photo matching (Kemp et al. 1997; Bruce et al. 1999; Burton, White & McNeil, 2010). Performance on all of these tasks has been found to be highly error prone, with people making around 10-20% errors in identity judgments when matching faces, depending on the details of the task (Burton et al. 2010, Bruce et al. 1999 & Kemp et al. 1997). Furthermore, experience and training in face recognition appears to make no difference to task performance - passport officers performed no better than a random sample of undergraduate students in a recent face identity task that mimicked the passport control face-matching procedure (White et al. 2014).

The fact that people are making around 1 in 5 errors in such tasks is particularly worrying, given that this performance level is from tasks where people (or photograph face stimuli) viewed in the tasks, have been cooperating with the procedure, that is, they have not been trying to subvert the identification (Henderson, Bruce & Burton, 2001; Burton et al. 2010). The stimuli for face recognition experiments are often new, controlled photographs taken specifically for the study. For instance, face photographs are usually taken under consistent lighting, using good quality cameras, posing a neutral expression and captured from a front on angle. When several different images of the same person have been taken for a study, there has been no deliberate attempt to make images of the same person look different across these multiple photographs. Same person photographs have mostly been captured within very short time intervals – often only minutes apart – reducing natural variance in appearance due to change in style or age. It is thus very possible that performance on such tasks reflects a level of performance that is higher than would be achieved in less favourable but more realistic conditions that include incidental image variation. For example, current legislation allows a passport photograph to have been taken up to ten years prior to use. Matching across this decade span is likely

far more challenging than an experimental task for which all photographs were taken just several weeks before testing (White et al. 2014). Similarly, the image quality available from CCTV could be very poor compared with images in laboratory studies.

Another characteristic of previous face-matching studies is that they have drawn their mismatch identities from a rather limited pool. Different person trials have typically been created by pairing together people whose face images look a bit similar to each other – perhaps due to similar hair colour or face shape. All false match images have to be selected from the available pool of photographs in the stimulus set. Images used in mismatch trials are thereby not always convincingly similar in appearance, even for unfamiliar viewers. This could make them easy to reject.

The upshot is that face-matching ability for very similar faces could be even worse than in previous studies. In certain applied situations, for example, when using a fraudulent passport, people may have very strong incentives to use the identity of a person who looks naturally similar to them in appearance. Yet little is known about viewers' ability to discriminate highly similar faces.

**Familiarity**

Familiarity has been shown in past studies to predict performance accuracy for both same person and different person identification (Burton et al. 1999, Jenkins et al. 2011). Given the influence of familiarity on past studies of face recognition, familiarity may be an important factor in face-matching where tasks include naturalistic same person image pairs which encompass natural within person variation, and also different image pairs which are of extremely similar appearance.

Experimental participants regularly perform at ceiling level in face-matching tasks when the images available for comparison are of faces that are familiar to them (Hancock, Bruce & Burton, 2000). For example, highly accurate performance is achieved even when the image shown at testing differs from that shown at initial presentation in facial expression or in the photographed angle (Bruce, 1982). Jenkins et al. (2011) demonstrated that the familiarity effect holds strong when many photographs are compared and when the images available for comparison are uncontrolled, and highly varied on conditions including pose, lighting, expression, hairstyle, and age. In one of their studies, participants were presented with 40 shuffled face picture cards. Unknown to the participants, this card deck comprised of 20 face picture cards of one female Dutch celebrity, and the other 20 face picture cards were of another female Dutch celebrity. Both unfamiliar viewers (20 British participants) and familiar viewers (20 Dutch participants) were asked to sort the card deck by identity, grouping together photographs which they believed showed the same individual's face. Participants were given no time restriction and could make as many or as view groupings as they felt reflected the number of identities present. A strong familiarity effect was found in this experiment. Unfamiliar viewers struggled with the task, dividing the deck into 7.5 identity piles on average. But familiar viewers easily group the cards into the two correct identities. Familiarity with the faces concerned made the task easy, even though the stimuli were highly variable and this familiarity advantage extends to poor quality images (Bruce et al. 2001, Burton et al. 1999).

Findings such as these highlight that familiarity can make face identity decisions easy, even when the same decisions are difficult for unfamiliar viewers. However, there has been very little research into whether familiarity helps with distinguishing extremely similar faces, such as the face of a disguised imposter from the true identity. Some support for the notion that familiarity could help comes from studies on telling twins apart. Stevenage (1998) found that after corrective feedback training, participants rated photographs of the same twin to be more similar, and images of different twins as less similar. Robbins & McKone (2003) also report training participants to distinguish identical twins, also using corrective feedback to aid learning. This study focussed primarily on

holistic processing rather than identity judgment. Although these studies did not test the familiarity effect directly and in isolation (as they gave corrective feedback between trials) the finding that identical twin faces can be learnt and distinguished after much exposure to them during the training phase provides scope for a potential advantage for familiarity. As familiarity helps in the case of distinguishing the faces of identical twins, there is reason to think that familiarity with a face provides a good starting point for investigating performance in a matching task involving unrelated, but very similar faces.

**Familiarity: a Graded Effect**

There are problems in defining familiarity with a face. Experimentally, face groups have often been divided into familiar and unfamiliar faces for all participants – for example celebrity faces as the familiar face set (Clutterbuck & Johnston, 2002) and a convenience sample of non-celebrity faces as the unfamiliar set (Burton et al. 2010) or faces of celebrities from other countries who are unfamiliar to the participants being tested (Burton, Kramer & Ritchie, 2015). An alternative approach has been to create a familiar face set based on colleagues or classmates of the experimental participants and to compare their performance with another group of participants who would be unlikely to know the target faces (Burton et al. 1999). Still other studies have familiarised viewers with novel faces as part of the experiment (Clutterbuck & Johnston, 2004, 2005). These methods are not without their faults. For example, not all faces that are presented as familiar (usually celebrity faces) are familiar to all participants. Even among the faces that are known to the participants, it is unlikely they will be equally familiar to them. For these reasons, a face cannot always be neatly categorized as familiar or unfamiliar to a single participant, let alone to a group of participants.

Clutterbuck & Johnston were among the very first to demonstrate the graded nature of familiarity experimentally (Clutterbuck & Johston, 2002; 2004; 2005). In the first paper in this series, the familiar and moderately familiar faces used in the study were all celebrity face images, with each celebrity's familiarity category chosen on the basis of familiarity

ratings provided by eight independent raters. An increase in familiarity led to a decrease in the time taken to match a full-face image to images showing just the internal features of a face (see Figure 2.1). Participants were fastest at matching a picture of the full face to images of internal features for highly familiar faces, slower for moderately familiar faces, and slowest for unfamiliar faces (Clutterbuck & Johnston, 2002). There was however no significant difference in performance when matching full-face images of each of the categories (highly familiar, moderately familiar, unfamiliar) to same or different images of the external features of the face.

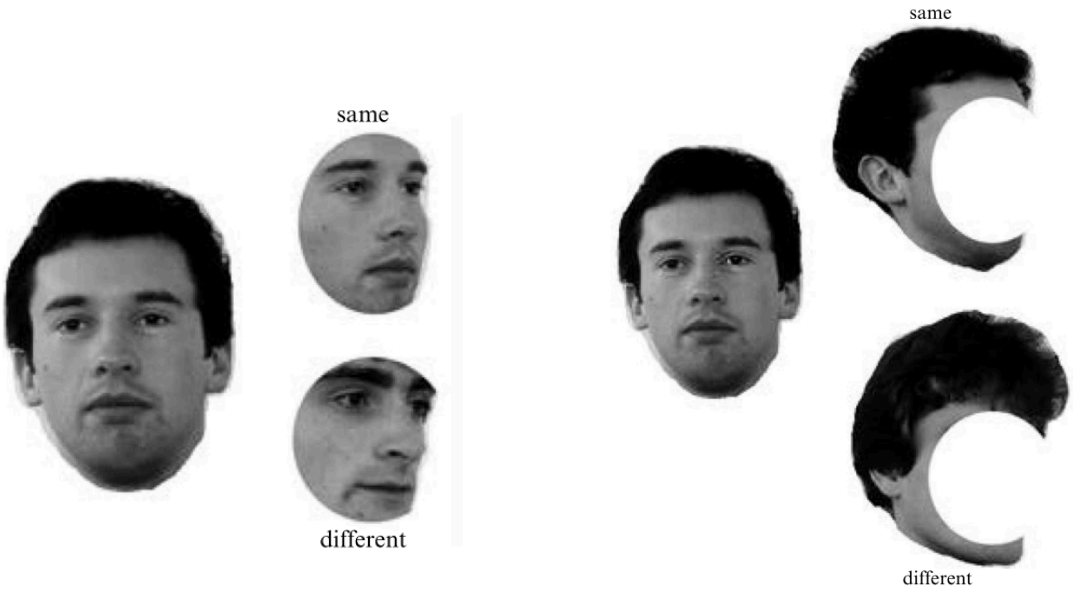


**Figure 2.1** Example of full face and internal feature stimuli (left) and full face and external features (right) viewed as part of Clutterbuck & Johnston's (2002) face-matching task.

Clutterbuck and Johnston (2002) assumed that the faces in each of the categories were of the same familiarity level for all participants (i.e. a face in the highly familiar category was assumed to be highly familiar for all participants). However, the raters who rated the familiarity of the face, did not take part in the experiment itself, so the familiarity bands may not be an accurate reflection of how familiar each face was to each of the participants in the main study. It is possible that some participants would be more familiar with the moderately familiar celebrities and *vice versa*.

Clutterbuck and Johnston (2004; 2005) carried out two additional studies that address these concerns to some extent. They used celebrity face images for the familiar category, newly learnt faces for mid-level familiarity and novel (previously unseen) faces to provide the unfamiliar level. The graded familiarity advantage was also observed using this method of familiarity division. Performance improved as familiarity increased on tasks involving gender judgement (Clutterbuck & Johnston, 2004) and face-matching speed (Clutterbuck & Johnston, 2005). Taken together, these findings strongly support familiarity as a continuous variable rather than a binary concept – people can be more or less familiar with a face, and this level of familiarity will affect performance on face-matching tasks in a graded way.

When investigating the effect of familiarity on face-matching ability for very similar faces, and using images of same faces that incorporate natural variation, it will be important to ensure that familiarity is measured in a way that accounts for i) familiarity being a graded concept and ii) the notion that not all celebrity faces will be equally familiar to all viewers.

## 2.3 Celebrity Faces & Celebrity Lookalikes

Testing for a familiarity advantage requires faces that differ in their degree of familiarity. Most people are highly familiar with their friends, colleagues and family members. The problem with this is that familiarity with these faces is very specific to the few individuals who know them. A study on face familiarity requires a large number of images of faces that will be familiar to many people who view them in the study, and also a large number of images of faces that are unfamiliar to these people. Ideally each of the faces used across the experiment will be familiar to some participants but unfamiliar to others, and all of the faces will be familiar to at least some of the participants. Celebrities provide a group of identities who are familiar to a very large array of people, and celebrity face images are easily accessible via Internet search. An additional advantage of using celebrity images is that there are many different categories of celebrities including pop stars, reality television stars, actors, politicians and sports personalities. This range gives the

freedom to choose images of celebrities from very different settings. Based on personal interests, each participant will be more or less familiar with celebrities from each of these categories. If celebrity images are sampled from a wide range of categories (singers, actors, politicians etc.), it is likely that each participant will be familiar with at least some of these celebrity faces and less familiar with others.

**Use of Lookalikes as a Proxy for Imposters**

In order to investigate performance for a challenging face-matching task proposed, it is necessary to have access not only to different photos of the same face that vary naturally in their appearance - such images are known as *ambient* images (Jenkins et al. 2011; Sutherland et al. 2013) - but also to photos of other faces that look extremely similar to the targets.

Conveniently, there is a ready source of faces that closely resemble celebrities and that is the celebrity lookalike industry. In the following experiments, I will use celebrities and their professional lookalikes to construct highly similar face pairs.

## 2.4 Experiment 1: Lookalike Task

The purpose of Experiment 1 was to test the effect of familiarity on performance accuracy in a challenging face-matching task. All of the face stimuli are ambient images to sample the natural variation in each person's appearance. This will presumably result in more challenging same person trials than in past work that has used highly controlled cooperative images. The use of celebrity lookalikes as imposters in my experiment should allow for extremely difficult different-identity pairs, compared with those used in previous experiments. The intention here is to model real life situations where someone may be trying to pass impersonate a similar looking person on a fraudulent security document.

If familiarity helps with these very fine distinctions, then matching performance should be more accurate for increasing levels of familiarity.

## Method

### Participants

30 undergraduate students (M = 8, mean age = 20.2) at the University of York volunteered as participants for this project. All participants were paid £3 or a half hour course credit in return for their participation.

### Stimuli

#### Image Selection

Face images of 30 celebrity identities (three face images per celebrity), and one face image of a professional lookalike for each of these celebrities, were selected as experimental stimuli (120 images in total). This number of images was necessary to provide two celebrity face photographs to constitute the same-person pair, and one additional celebrity face image to be paired with the lookalike image to create the different pair in the face-matching task. For a celebrity to be included, they needed to have at least one professional lookalike whose image was accessible from the Internet. For a lookalike image to be chosen as suitable for use in the study, a viewer who was extremely familiar with the celebrity in question approved their high level of visual similarity to the celebrity.

Following approval of the celebrity and lookalike images in the image selection, two face image pairings were created for each celebrity – one showing two different images of the celebrity, and the other showing a third image of the celebrity paired with an image of that celebrity's lookalike (see Figure 2.2). The lookalike image could appear on either side

of the screen, and appeared on the left and right side equally often across the experiment.


**Figure 2.2** Face-matching task image examples. The pairs on the left (A) show different identities (with the imposter face on the right), the pairs on the left (B) show same identity pairs.

**Design**

This experiment was conducted using a within subjects design, which tested the effect of five levels of face familiarity on the dependent variable, which was performance accuracy in the face-matching task.

A novel familiarity scale was designed for use at the end of the experiment to assess how familiar each of the celebrities was to each of the participants. This scale was set across the desk where the experiment took place and was a meter long in length, marked for 0-100cm, with 0 representing unfamiliar faces and 100 representing faces that were extremely familiar (see Figure 2.3). This scale was intended to address two limitations of previous familiarity manipulations, i) not all familiar (celebrity) faces are familiar to everyone, ii) some faces will be better known than others and this is the case for each participant.

## Procedure

### Face-Matching Task

Participants took part in a face-matching task involving 60 image pairs (two pairs for each of the 30 celebrity identities), viewed on a computer screen. The participants' task for each pair was to determine whether the two images showed the same identity or different identities (i.e. one of the images was of the lookalike). Two different random orders of image pair presentations were created; each participant was assigned to view one of these two random orders. Participants were informed that the lookalike images could appear on either side of the screen and that there was no time limit for completing the task.

### Familiarity-Rating

On completion of the face-matching task participants were given photograph cards (size 6cmx4cm) of each of the celebrities that they had seen in the face-matching task (N = 30) and asked to rate them for how familiar they were with the celebrity's face before completing the task. Participants received just one of the three true celebrity images that they had viewed in the face-matching task for use in the familiarity-rating task. The image viewed for each celebrity was selected randomly from the three available, and the chosen card for each celebrity remained the same for all participants. Participants rated the faces for familiarity by placing them on a scale that ran from 0 (completely unfamiliar) to 100 (extremely familiar). Faces of equal familiarity could be placed down vertically one above the other to create a column of equally familiar faces; this was particularly useful for faces that were completely unfamiliar and extremely familiar (see Figure 2.3 for image of the familiarity scale in the experimental setting). This approach allowed me to capture the relative and absolute familiarity of the celebrities separately for each participant.

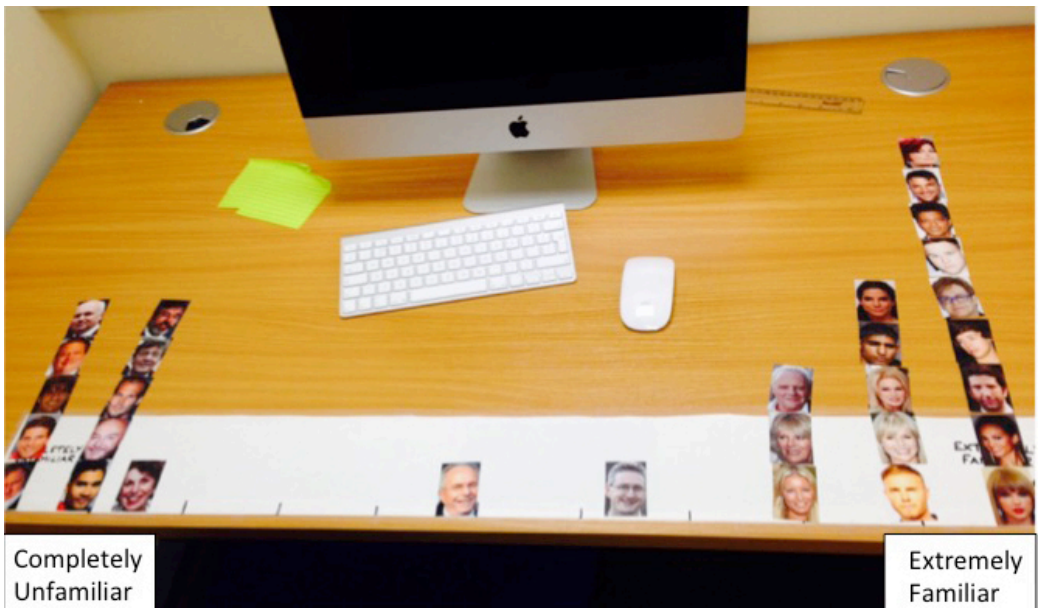


**Figure 2.3** Photograph of one participant's use of the familiarity scale taken immediately after completion in the experimental setting. The far left side of the scale indicates that the face was completely unfamiliar, and the far right depicts extreme familiarity with the face.

### Analysis

The main measure of interest was the percentage of correct responses in the face-matching task. To examine the effect that familiarity score had on percentage accuracy in the face-matching task, I grouped the raw familiarity ratings into 5 different familiarity levels. This was achieved by binning each participant's face matching responses into 5 familiarity bands (quintiles) based on the participant's own ratings. In total 35% of faces were placed in Band 1, 9% in Band 2, 10% in Band 3, 8 % in Band 4 and 39% in Band 5. Interestingly, although the majority of faces were placed at the extremes of the spectrum (0-19 and 80-100) all participants placed some faces in the middle familiarity bands (20-39,40-59, 60-79).

### Results

Percentage accuracy scores were entered into a one-way repeated measures ANOVA to investigate the effect of familiarity on face-matching performance in this celebrity versus lookalike discrimination task. Percentage scores were rounded to integers throughout.

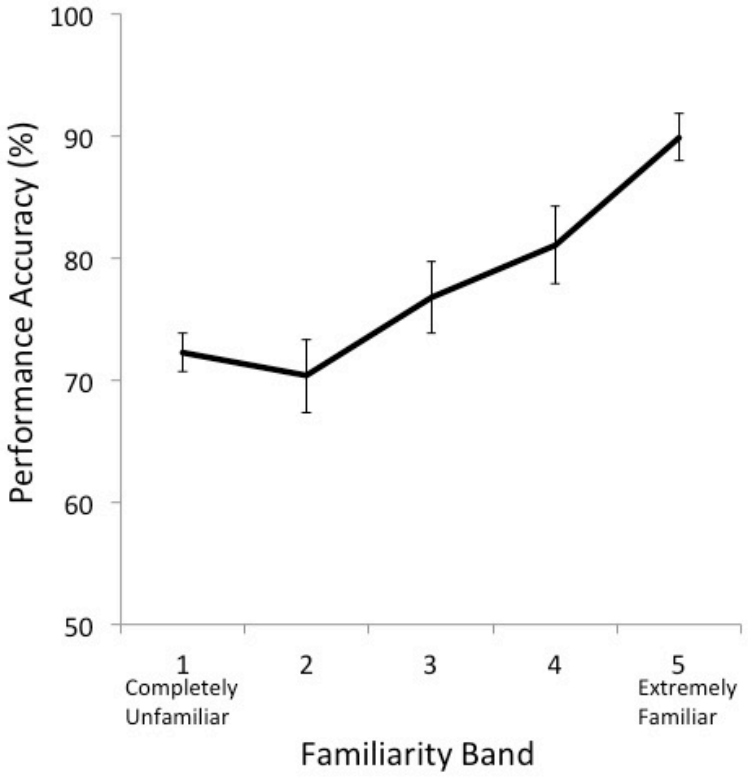


**Figure 2.4** Percentage of correct responses in face-matching task (using fine quality 200x300 pixel images) for each familiarity quintile; 1 (0-19), 2 (20-39), 3 (40-59), 4 (60-79), 5 (80-100). With Band 1 being completely unfamiliar and Band 5 being extremely familiar. Error bars show standard error of the mean.

The results revealed that face-matching performance was significantly affected by the *Familiarity* with the face viewed, F(2.98, 86.51) = 10.07, p = <.001, $\eta_p^2$ = .26 (Greenhouse-Geisser corrected). Accuracy was lowest for the faces that were most *Unfamiliar* (Band 1, M = 72%, SE = 1.58, CI = 69 -76), and best for the faces that were most *Familiar* (Band 5, M = 90%, SE = 1.95, CI = 86-94). As can be seen in Figure 2.4, there was a generally graded increase in performance as familiarity increased (Band 2: M = 70%, SE = 2.98, CI = 64 -76; Band 3: M = 77%, SE = 2.94, CI = 75-88; Band 4: M = 81%, SE =1.95, CI = 77-84).

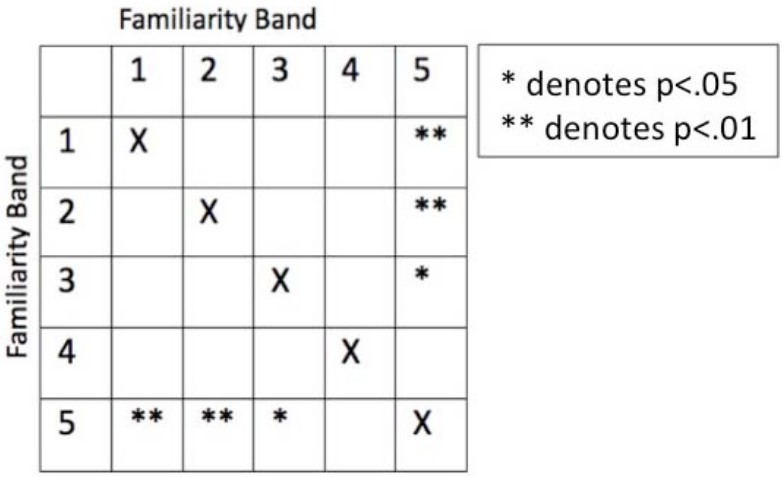


**Figure 2.5** Pairwise comparisons showing which familiarity levels performance was significantly better than the other familiarity levels.

Pairwise comparisons revealed that the performance was significantly better for faces that were placed in familiarity Band 5 than for faces in Bands 1,2 and 3 but not 4 (see Figure 2.5).

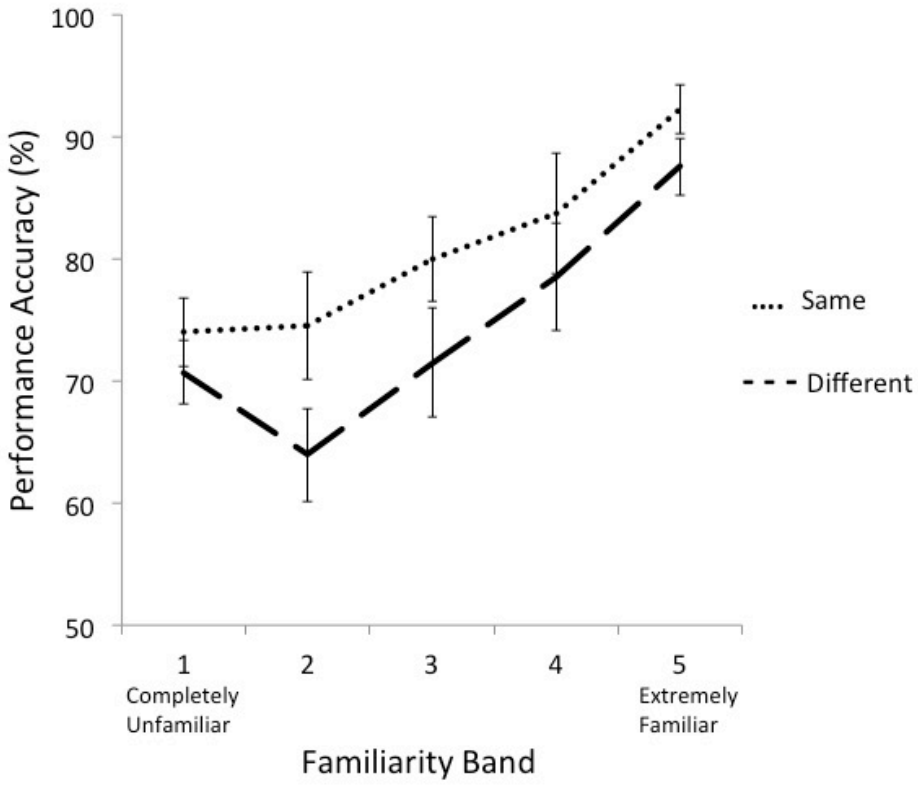


**Figure 2.6** Graph showing pattern for *Same* identity pairs correct response and *Different* (lookalike) identity pairs correct responses. Error bars show standard error of the mean.

I then analysed accuracy separately for *Same* identity and *Different* identity trials (Figure 2.6). For both types of trial accuracy increased as familiarity increased. As expected there was somewhat higher accuracy in the *Same* identity condition than in the *Different* identity condition, presumably reflecting a tendency to judge highly similar faces as the same person. A 5x2 within-subjects ANOVA with the factors of *Familiarity* and *Pair Type* confirmed that this difference was significant: overall accuracy was significantly higher for *Same* trials (M = 81%, SE = 2.11, CI = 77-85) than for *Different* trials (M = 74%, SE = 1.84, CI = 71-78), [$F(1,29)=5.31$, $p<.05$, $\eta_p^2=.23$). As expected, there was a main effect of familiarity [$F(4,116)=11.10$, $p<.001$, $\eta_p^2=.12$], indicating a graded familiarity advantage for *Same* and *Different* identity trials alike. There was no interaction between *Familiarity* and *Pair Type* – familiarity had no more of an effect for correct performance on *Same* identity trials than *Different* identity trials [$F(4,116)=.38$, $p=.83$, $\eta_p^2=.01$].

**Discussion**

Participants were able to make remarkably fine discriminations between extremely similar faces (celebrities and their lookalikes), and to integrate naturally varied same identity images. This ability was underpinned by a graded effect of familiarity. Accuracy was at its peak when the faces viewed were extremely familiar to the participants (performance accuracy 90%), yet performance was much worse for the completely unfamiliar faces (accuracy 72%, note that chance performance is 50%). Task performance generally increased with a progression in familiarity (Band 2 M = 70%, Band 3 M = 77%, Band 4 M = 81%). Importantly these results are based on personal familiarity scores for each of the celebrities. This method of analysis allowed a graded familiarity effect to be teased out despite the fact that not all participants were familiar with the same celebrities.

Performance for completely unfamiliar faces in the lookalike study (72% accuracy) was lower than identity judgment accuracy found in past studies (e.g. Bruce et al, 1999; Megreya & Burton, 2007; Burton et al. 2010). When compared to performance on the GFMT, a standardised test of face-matching ability that comprised of cooperative stimuli and limited false match image options, mean performance accuracy is lower on the lookalikes task. This is the case for overall performance accuracy on the unfamiliar face images (lookalikes task mean performance =72%, GFMT = 89.9% (long version), 81.3% (short version), and also for same identity trials (lookalikes = 74%, GFMT = 92% [long version], 79.8% [short version]) and different identity trials (lookalikes = 70.7%, GFMT = 88% [long version], 82% [short version]). Strikingly, performance is nearly 20% worse for the naturally varied images of the same celebrity faces in my study, than the cooperative same person images used in the GFMT. Furthermore, performance is more than 10% worse for different person trials in my lookalikes task than even in the short version of the GFMT, which includes only the hardest items in the GFMT. This highlights that both photographs of the same person which include natural variation, and very similar looking different identities, cause more problems to face matching performance than previously

captured by tasks constructed from cooperative stimuli. This is concerning given that both of these image types may be encountered in security situations and attempts of fraud.

As in past face recognition research which has shown a familiarity advantage for both face memory and face-matching (Bruce, 1986, Ellis et al. 1979, Burton et al. 1999), familiarity had a great impact on face-matching performance. The present study shows that the benefit of familiarity extends to the challenging case of extremely similar lookalike faces, and naturally varying images. Whereas past studies were often very easy for familiar viewers, resulting in ceiling performance (Hancock, Bruce & Burton, 2000), the lookalike task brought performance off ceiling, so that modulations in performance could be observed. Even with these very challenging viewing conditions accuracy reached 90% in the highest familiarity band. Thus, it seems that the effect of familiarity is so strong, that even professional lookalikes are an unconvincing false match for viewers who are extremely familiar with that celebrity.

My study broke down familiarity even further than in previous studies, by comparing 5 levels of familiarity rather than just 3 as used by Clutterbuck & Johnston (2002, 2004, 2005). Assessing performance across 5 levels of familiarity provided a more realistic and accurate categorisation of participants' familiarity with the faces, allowing a more detailed exploration of the extent of graded effect of familiarity. Familiarity with a face is not an all-or-nothing phenomenon. Different viewers are familiar with different faces to differing degrees. Constructing the analysis around that insight reveals much finer structure than a binary familiar/unfamiliar distinction allows.

Now I have demonstrated that the graded familiarity advantage survives to the challenging case of imposter detection, I want to see how much further I can push the effect of familiarity. So far, I have only tested the effect for fine quality (un-manipulated) images. Although I have found that familiar viewers can tell apart a target face and imposter, it is not clear what they have learnt about the familiar face which has allowed

them to do this. Given that the faces in the task were very similar (i.e. celebrities and their lookalikes), it seems likely that fine scale information in the image is critical for making the necessary discriminations. If so, then obscuring the fine-scale information should impair performance, even for familiar viewers. There are several possible techniques for obscuring image detail. Here I used image pixelation, whereby the number of pixels in the image is reduced while image size is held constant. This technique has an interesting pedigree in the psychology literature (Harmon & Julesz, 1973). It also arises in the context of applied face identification whenever digital images are enlarged (Jenkins & Kerr, 2013).

## 2.5 Experiment 2: Mid Pixelation

In Experiment 1 a graded effect of familiarity on face matching performance was evident for both integrating different images of the same face and telling apart very similar faces of different identities. In that experiment, fine quality images were used as stimuli. However, face images encountered in applied matching tasks are often not of good quality. For example, coarsely pixelated images may be obtained by zooming in on Closed Circuit Television (CCTV) footage to gain an image of a suspect's face (see Figure 2.7 for applied example).

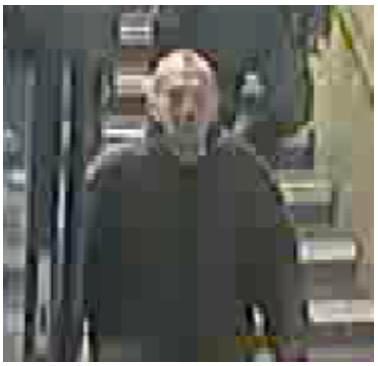

**Figure 2.7** Example of actual image issued by the police to the public to assist with identification of a man caught on CCTV (Howarth, 2016). This image takes a pixelated appearance.

Face recognition for unfamiliar faces is poor even under favourable image conditions (see Burton et al. 2010). The task is even more challenging when the images are of low

resolution as pixelation disrupts the information that is available from an image (Harmon & Julesz, 1973). When an image is pixelated, the information from several adjacent pixels is pooled to form larger pixels. The luminance of the new pixel is determined by averaging together the luminance values of the constitute pixels. High spatial frequency information within that area is thus lost. In tandem, high spatial frequency noise is introduced at the edges of the new pixels. This is due to the larger changes in luminance between adjacent pixels in the new image than between adjacent pixels in the original image. This noise is particularly disruptive to viewing, as the visual system is highly sensitive to lines and geometric patterns, making the new pixel boundaries difficult to ignore (Harmon & Julesz, 1973).

As a result of this manipulation, it is difficult to extract exact information about a face from a pixelated image. In particular, when we view a pixelated face image, configural information (the metric distances between features) that can be extracted from a face becomes less precise, and the appearance of features becomes less detailed. Unsurprisingly, the pixelation manipulation has been shown to increase the difficulty of image recognition, compared to a non-manipulated version of the image (e.g. Harmon & Julesz, 1973, and replicated by Morrone, Burr & Ross, 1983; Sergent, 1986; Bachmann, 1991; Costen, Parker, & Craw, 1994, 1996; Uttal Baruch & Allen, 1995). However, people can *name* pixelated celebrity images, even at very low levels of pixelation (Lander, Bruce, & Hill, 2001). The work of Lander and colleagues (2001) focused on pixelation from the perspective of identity protection issues – images are often pixelated by the media in order to try and *protect* identity. They argued that pixelation is not an effective method of obscuring identity, as the face remains identifiable to a familiar viewer. Although *recognition* ability (being able to say who the face is) for pixelated images of famous or learnt faces has been tested in the past, face-matching (determining whether pairs of images depict the same person or different people) with pixelated faces has received little attention.

Bindemann, Attard, Leach & Johnston (2013) were the first to test performance in a matching task using pixelated unfamiliar faces. In their study participants were tasked with matching a pixelated image to a good quality image. The stimuli used in this study were the images used in the GFMT, where one image in each of the image pairs had been pixelated (see Figure 2.8). They found that participants were much better at matching two intact face images, than matching one intact face image to a pixelated image. Moreover, performance declined as the level of pixelation increased. However, in some practical situations investigators may have to compare multiple pixelated images from different CCTV footage to try and piece together an event sequence. I have been unable to find any published research that previously tested such performance. In addition to this, Bindemann et al. (2013) used staged face photographs, using either posed profile or front view face image (Burton et al. 2010). For example, images captured from CCTV footage may vary greatly in lighting and pose. It is unlikely that such photographs would be available in forensic settings. For these reasons, it is important to measure face-matching ability for pixelated images using ambient images, as past research has shown that the same face can appear drastically different between ambient photograph images (Jenkins et al. 2011).

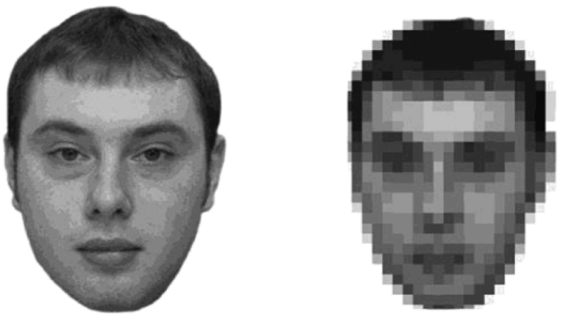

**Figure 2.8** Example of stimuli used in the face-matching task created by Bindemann et al. (2013)

These issues will be addressed by Experiment 2, which will test people's ability to match pairs of pixelated ambient images. This new task is more difficult than previous recognition and naming tasks because both the lookalike and the celebrity image map

onto the same individual. For example, the viewer might associate both images with say, Al Gore, but that is not enough to solve this task. The problem is to decide whether both images actually show Al Gore, or whether one of the images shows an imposter (lookalike). Additionally, the nature of the paired matching task gives a baseline score. Unlike in a naming task, chance performance is known to be 50% in the paired matching task, so observed performance can be compared against this chance level. Finally, the ambient images used in this task provide us with a test more similar to the image type available in real world investigations.

To address all of these issues I repeated Experiment 1, but this time replacing the fine quality ambient images with pixelated versions of these images. The aim is to establish whether the graded familiarity advantage observed for imposter detection in fine quality images extend to degraded images. If it is knowledge of exact facial configurations and fine featural detail that differentiates familiar and unfamiliar viewers in the lookalike task, then obfuscating that information should eliminate the familiarity advantage.

If familiar viewers are using the detail and small differences in faces to solve the task, then there should be little or no benefit of familiarity in face-matching performance in Experiment 2, where this information is difficult to access. On the other hand, if the familiarity advantage survives, that would suggest that other information is being used.

**Method**

**Participants**

30 undergraduate students at the University of York (M = 11, mean age = 19.7) volunteered as participants for this project. Participants received payment of £3 or a half hour course credit. None of the participants had taken part in the previous experiment.

**Design & Stimuli**

As in Experiment 1, a within-subjects design was adopted to compare the effect of *Familiarity* (5 levels) on the dependent variable, percentage of correct responses in the face-matching task.

The image pairings of the 30 celebrity faces and their 30 lookalikes faces were the same as Experiment 1. However, unlike Experiment 1, all of these images were pixelated to a level of 30 pixels wide 45 pixels high using Adobe Photoshop (CS6) (see Figure 2.9 for a side by side example of the stimuli used in Experiment 1 and the pixelated versions for use in Experiment 2).

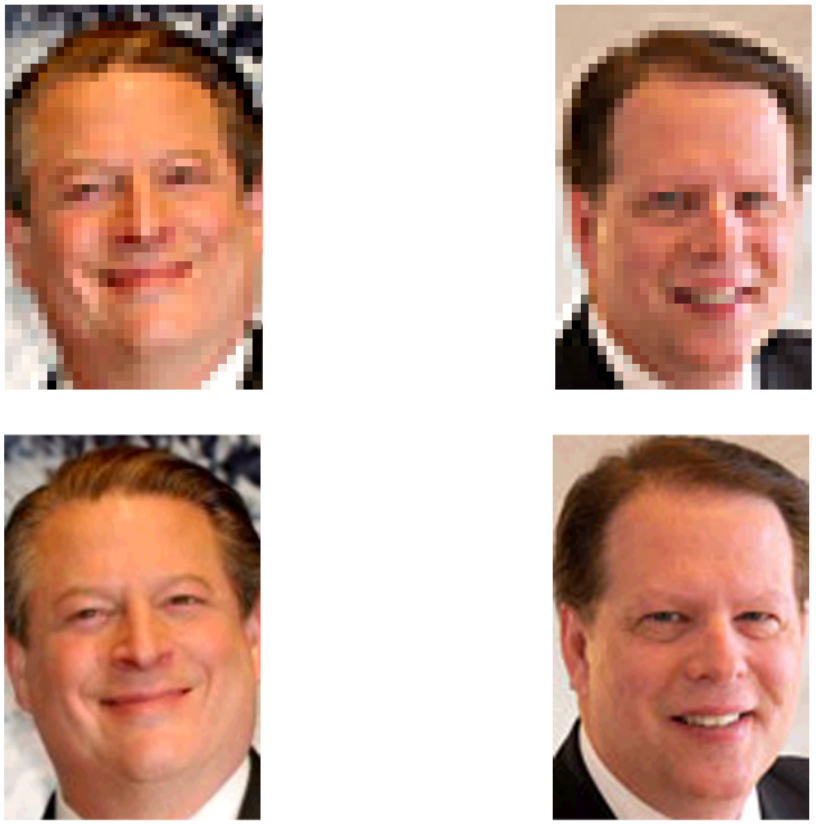

**Figure 2.9** Example of the image appearance for Experiment 2 (top pair) compared with the fine version of the same image as used in Experiment 1 (bottom) pair. These are different image pairs of Al Gore with the lookalike appearing on the right.

**Procedure**

The procedure was the same as in Experiment 1, except that this time participants viewed the newly created pixelated versions of the image pairs.

Following completion of the face-matching task, participants ranked the celebrity faces for familiarity using the familiarity scale as in Experiment 1. Full resolution (un-manipulated) images were used for the familiarity-ranking task as in Experiment 1.

**Analysis**

As in the previous experiment, results were analysed by comparing the percentage of correct responses across the familiarity quintiles. In this experiment 34% of faces were placed in Band 1, 12% in Band 2, 9% in Band 3, 10% in Band 4 and 34% in Band 5.

**Results**

A one-way repeated measures ANOVA was performed on the accuracy data, to investigate the effect of familiarity (5 levels) on participants' face-matching ability for poor quality images.

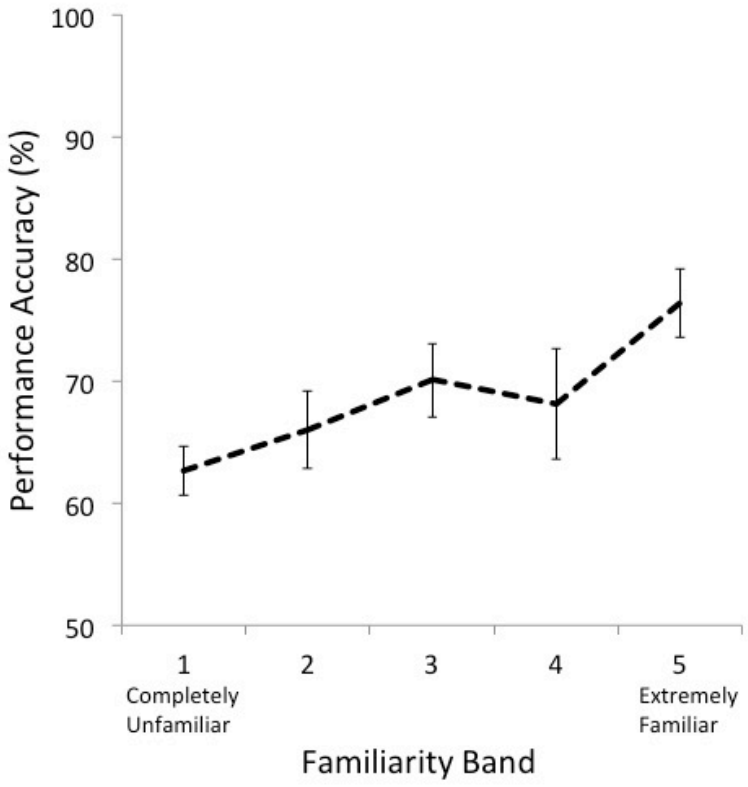


**Figure 2.10** Graph showing the graded effect of familiarity for participants' face-matching task performance. Error bars show standard error of the mean.

As in Experiment 1, face-matching performance was significantly affected by *Familiarity* with the face viewed, F(3.19, 92.48) = 2.96, p<.05, $\eta_p^2$ = .09 (Green-House Geisser corrected). Figure 2.10 shows the predominantly graded effect of *Familiarity* over the 5 familiarity bands (Band 1: M = 64, SE = 2.08, CI = 59 -68; Band 2: M = 66, SE = 3.22, CI = 59-73; Band 3: M = 70, SE = 2.96, CI = 64-76; Band 4: M = 68.69, SE = 4.07, CI = 60-77; Band 5: M = 76, SE=2.79, CI = 70-82).

Pairwise comparisons revealed that accuracy for the highly familiar faces (Band 5) was significantly better than performance for faces in familiarity Band 1, mean difference = 12.28, SE = 2.75, CI = 6.65-17.91, p<.005 and also than that of Band 2, mean difference = 9.75, SE = 3.96, CI = 1.64-17.85, p<.05. There were no significant differences in

performance accuracy between each of the other familiarity bands (p>.05 for all comparisons).

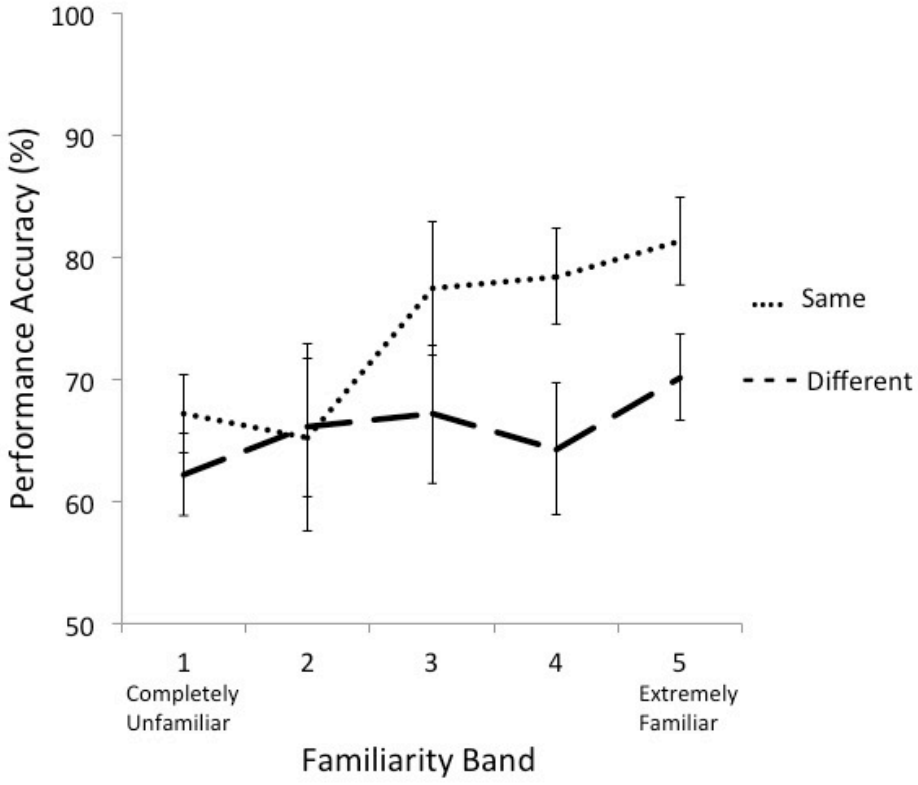


**Figure 2.11** Face-matching performance broken down into correct *Same* and *Different* identity trials. Error bars show standard error of the mean.

Accuracy data were also analysed according to the breakdown of *Same* identity and *Different* identity correct trials using a 2x5 ANOVA (see Figure 2.11). A significant main effect of *Familiarity* was observed [F(4,116) = 3.00, p<.05, $\eta_p^2$=.09]. There was also a significant main effect of *Trial Type* - participants were overall more accurate at the *Same* identity trials than the *Different* identity trials [F(1,29) = 4.99, p<.05, $\eta_p^2$=.15]. However, there was no significant interaction between *Familiarity* and *Trial Type* [F(4,116)=.79, p=.53, $\eta_p^2$=.03] (see Figure 2.11).

**Discussion**

Despite the pixelation of the images, accuracy in the lookalike task improved as familiarity increased.

My pixelated lookalike task was naturally more challenging than previous identification tasks involving pixelated images. Firstly, the lookalike matching task is more challenging than naming pixelated celebrity images (Lander et al., 2001), as in my task both the lookalike and celebrity lookalikes could be mistaken for (and named as) being the same celebrity. In addition to this, previous matching tasks have involved matching one good quality image to one pixelated image (Bindemann et al. 2013), whereas my task required matching across two pixelated images. The findings from my challenging pixelated lookalike experiment hence demonstrated the versatility and strength of familiarity as an aid to face recognition, as even though the celebrity *and* the lookalike images were pixelated, the graded familiarity effect on face-matching performance prevailed. Finally, this study suggests that for familiar faces, learnt information other than featural details and configural information may be used to perform the task. Fine-scale image information was more difficult to access than in Experiment 1, yet the familiarity advantage survived despite this.

At the current level of pixelation (30x45 pixels) some featural and configural information is still visible. This raises the question of where the familiarity advantage will break down. Presumably in the limiting case (1x1 pixel), performance on this task would be at chance for all familiarity bands. The graded familiarity effect in the present experiment implies that viewers were nowhere near their performance limit in this task. In the next experiment, I set out to push the familiarity advantage to its limits by pixelating the images even further.

## 2.6 Experiment 3: Coarse Pixelation

Past research has shown that there is a limit to people's ability to recognise a pixelated face – we can recognize pixelated faces but only up to a point. The lower the image resolution the higher the error rate, with results eventually falling to chance level (Bachmann, 1991; Costen et al., 1994, 1996). It is believed that a familiar face can be recognised up to horizontal pixelation level of 16 pixels per face, any level beyond this results in a steep decline in performance (Bachmann, 1991; Costen et al., 1994, 1996). Bindemann et al. (2013) reported that the pixelation threshold was in fact much lower for matching unfamiliar faces. Bindemann et al.'s (2013) study reported limits on participants' abilities to match two side-by-side images, i.e. the scenario in the lookalike experiment series. Bindemann and colleagues (2013) reported that a large drop in face-matching performance accuracy occurred when one of the high-resolution images in each of the image pairs was replaced with images of a horizontal resolution of 20 pixels. However, Lander (2001) found that people could identify around half of the familiar face photographs presented to them when the images comprised of a horizontal resolution of only 10 pixels per face. In Bindemann et al.'s (2013) unfamiliar face-matching task performance was at around chance level for a horizontal pixelation of 8 pixels, even though the face-matching task consisted of co-operative stimuli (taken from the GFMT) presented side by side. Thus these previous findings indicate that pixelation will reach a point where the familiarity advantage no longer holds.

So far the graded familiarity effect has prevailed for ambient images of extremely similar faces, even when these images were degraded using pixelation. I previously argued that reducing the number of pixels in the image makes configural and featural information in the image more difficult to access. The accuracy data from Experiment 2 suggests that some critical information was still accessible albeit at a reduced level (which could explain the overall poorer performance in Experiment 2).

If the familiarity advantage remains for even coarser pixelation, it would suggest that information other than fine featural details and exact configurations support high performance by familiar views in the lookalike task. If the familiarity advantage is eliminated, this would suggest that the familiarity advantage relied solely on the fine-scale information that is disrupted by coarse pixelation.

## Method

### Participants

30 undergraduate students at the University of York (M = 8, mean age = 20.2) volunteered as participants for this project. Participants received payment of £3 or a half hour course credit. None of the participants had taken part in the previous experiment.

### Design

As in the previous experiments, Experiment 3 was a within-subjects study, which investigated the effect of familiarity on performance on the lookalike matching task. The only difference between this experiment and the preceding experiments was the level of pixelation in the stimulus images.

The pixelation level chosen for this experiment was 20 pixels wide x 30 pixels high. This particular resolution was selected because Bindemann et al. (2013) reported a marked drop in performance accuracy for this level of pixelation, although accuracy was still above chance in their study.

**Stimuli**

The same image pairings of the 30 celebrity faces and lookalike faces for each of these celebrities were used as in Experiment 1 and Experiment 2. This time the images were presented at a pixelation level of 20x30 pixels using Adobe Photoshop (CS6).

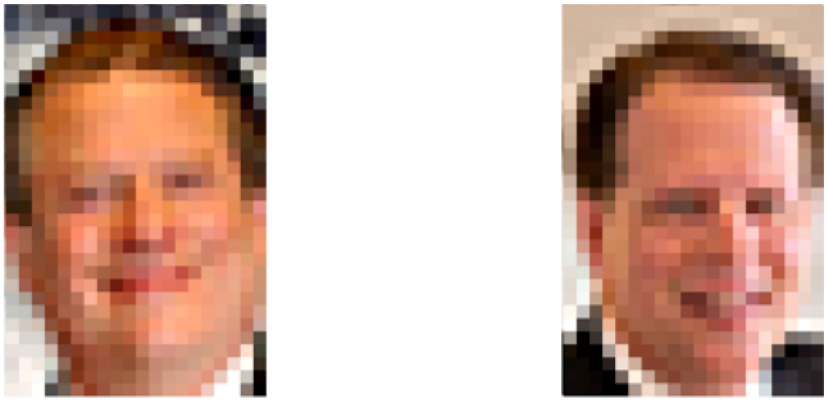

**Figure 2.12** Example of coarsely pixelated image stimuli used in Experiment 3.

**Procedure**

The procedure for this experiment was the same as for Experiments 1 and 2, except that the face-matching task now involved the coarsely pixelated images (20x30 pixels).

Following completion of the face-matching task participants used the familiarity scale as in the previous experiments, to indicate their level of familiarity with each of the celebrity faces. As in both previous experiments, the good quality image cards were used for the familiarity judgement task.

**Analysis & Results**

Experiments 1 and 2 established a graded effect of familiarity on task performance by examining accuracy at each of five familiarity bands. This approach was not possible for Experiment 3 because participants used the middle range of the familiarity scale less frequently. 37% of faces were placed in Band 1, 20% in Band 2 and 43% in Band 3.

Dividing the data into familiarity quintiles meant that the middle quintiles (2, 3 and 4) were too sparsely populated to allow meaningful statistical analysis. To circumvent this problem and obtain a reliable performance estimate for mid-level familiarity faces, data from familiarity bands 2, 3 and 4 were pooled into a single band. This resulted in 3 familiarity bands, as used in previous studies (e.g. Clutterbuck & Johnston, 2002, 2004, 2005).

A one way repeated measures ANOVA was performed on the accuracy data, this time examining the effect of *Familiarity* (3 levels), on performance accuracy in the face-matching task.

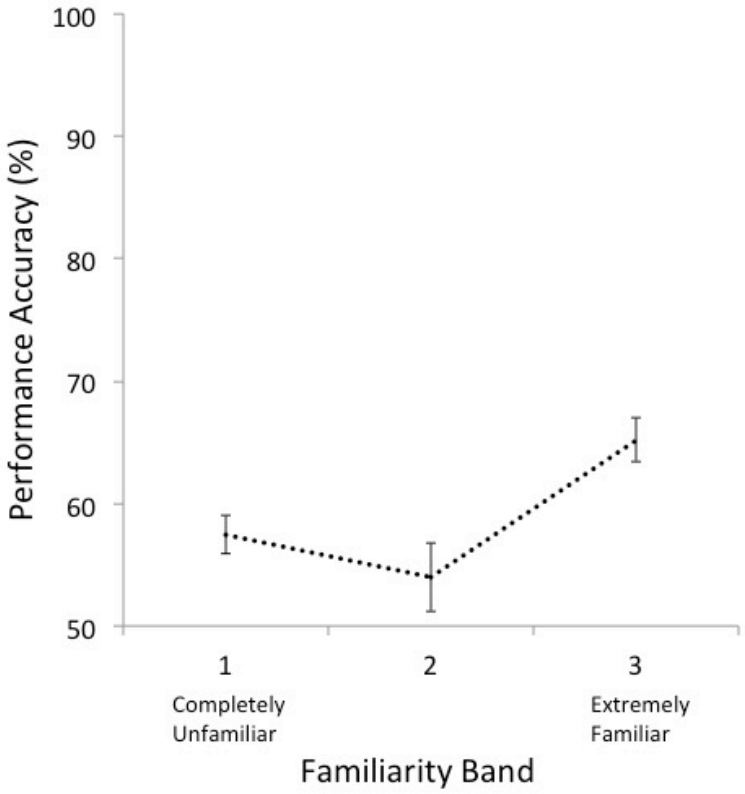


**Figure 2.13** Percentage of correct responses for each of the three levels of familiarity in the 20x30 pixel condition. Error bars show standard error of the mean.

Once again, there was an overall effect of *Familiarity* for this task [$F(2,58) = 9.44$, $p<.001$, $\eta_p^2 = .25$] (see Figure 2.13). Mean performance accuracy was highest for highly familiar

faces, Band 1 M =65%, (SE =1.78, CI =62-69). Performance accuracy for Band 2 was M = 54% (SE =2.84, CI = 48-60) and for Band 1 M = 58%, (SE =1.59, CI = 54-60) (see Figure 2.13).

Pairwise comparisons revealed that accuracy was significantly higher for highly familiar faces (Band 3) than for both faces of mid familiarity (Band 2) (mean difference = 11.22, CI = 4.67-17.77, p<.005) and unfamiliar faces (Band 1) (mean difference = 7.70, CI = 3.57-11.83, p<.005). There was no significant difference between accuracy scores for faces in familiarity Band 1 and Band 2.

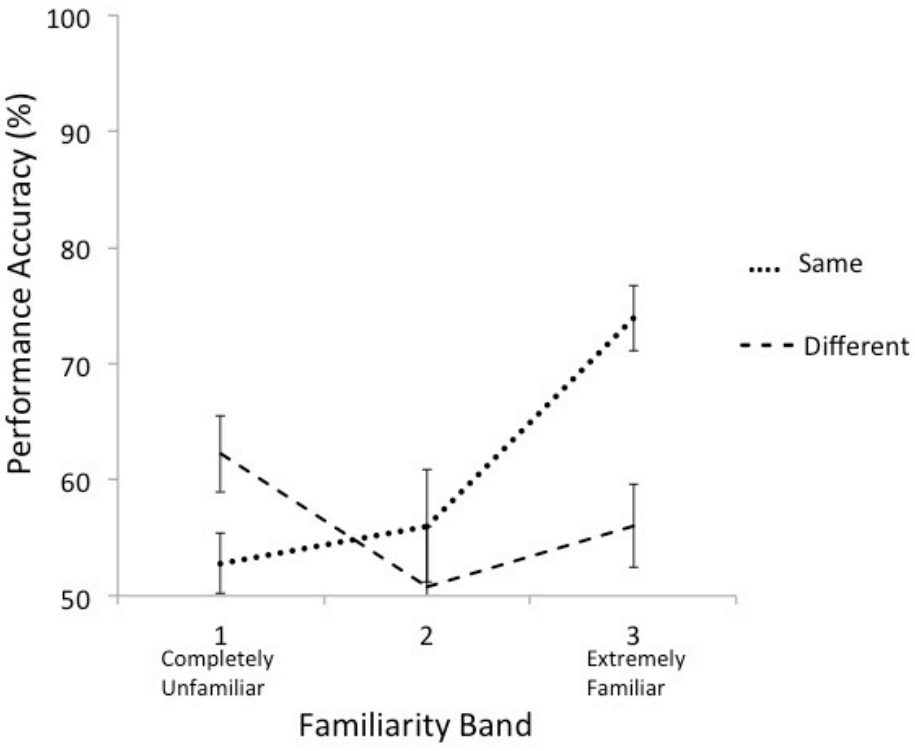


**Figure 2.14** Percentage of correct responses in face-matching task (using poor quality 20x30 pixel images) by familiarity broken down into same (dotted line) and different (dashed line) correct trials. Error bars show standard error of the mean.

A 2x3 ANOVA was conducted to break down results into *Same* identity and *Different* identity correct trials. This analysis revealed a significant main effect of *Familiarity* [F(2,58)=6.39, p<.005, $\eta_p^2$ =.18]. There was however no significant main effect of *Pair Type* [F(1,29)=1.09, p=.19, $\eta_p^2$ =.04]. There was also a significant interaction between *Pair Type* and *Familiarity* [F(2,58)=6.08, p<.005, $\eta_p^2$ =.17). This is illustrated in Figure 2.14.

There was a significant simple main effect of *Familiarity* for *Same* identity trials F(2,116) = 10.29, p<.001, $\eta_p^2$ = .15. There was no simple main effect of *Familiarity* for *Different* identity trials F(2,116) = 2.55, p>.05, $\eta_p^2$ = .04.

As a significant interaction was observed between *Familiarity* and *Pair Type*, Tukey post hoc tests were conducted to find out where significant differences lay. For *Same* identity trials, there was a significant difference in performance between familiarity bands *Low (Band 1)* and *High (Band 3)*, and between *Mid (Band 2)* and *High (Band 3)*. There were no significant differences between any of the levels of *Familiarity* for *Different* identity trials.

**Discussion**

Although a familiarity advantage survived in Experiment 3, with people being significantly better at recognizing the faces that were extremely familiar to them (M = 65% than the least familiar face (M = 58%), the graded effect seen in Experiments 1 and 2 did not emerge here, and accuracy in familiarity Bands 1 and 2 was numerically not much above chance level (50%). It seems that by reducing image quality to 20x30 pixels we are approaching the limit of the familiarity advantage in this situation. Breakdown of results by trial type revealed that this familiarity advantage was driven by improved performance for same person trials.

It is interesting that a familiarity advantage still emerged when comparing extremely familiar faces to less familiar faces. This advantage suggests a role for coarse scale information even when discriminating extremely similar faces. Performance was at around chance level unless the faces were extremely familiar. In some ways it may not be surprising that there was a performance advantage for highly familiar faces, as it has been previously shown that people can recognise associates even in very poor quality images (e.g. Burton et al. 1999). However, the foil faces in previous studies have been generic similar faces that merely share the same basic description (e.g. young male, short black hair). The important difference here is that my lookalike foils were themselves recognisable as the celebrities they were impersonating. The implication is that the lookalike faces differ from the celebrity faces only in subtle detail. Yet disrupting the subtle detail in the images was not catastrophic for familiar viewers. This suggests that the familiar viewers used other information to solve the task. One possibility is that at least some of the subtle differences are carried in the low spatial frequency information that is intact in the pixelated images.

Another cue comes from the pattern of breakdown of the familiarity advantage. In Experiment 3 familiarity did not improve performance on different identity pairs. It is thus possible that familiarity is making it easier to determine the identity of the target celebrity from the poor quality image, this identity decision could come from either the celebrity or lookalike image. When a familiar viewer can identify the celebrity, they then have access to all the representations that they have stored for that celebrity's face. Familiar viewers are aware of many more ways the celebrity's appearance can take, and hence allow a greater range of variation of appearances for the face, than an unfamiliar viewer may. Therefore familiar viewers may be more accepting of saying *same* to the image pairs in the matching task, even though the exact details can't be extracted from the pixelated images. This approach would improve performance for *same* identity trials but lead to poorer performance for *different* person trials, and could thus explain the pattern of results in Experiment 3.

It seems that faces are compared in different ways depending on our familiarity with the faces involved (e.g. Megreya and Burton, 2006). If people are unfamiliar with the face they may be matching face images in a pattern matching type manner, similar to the method used to match images of objects (Hancock et al. 2000; Burton & Jenkins, 2011), yet when people are extremely familiar with a face, our findings suggest that people no longer rely on this pattern type matching, but can use other information which we have learnt for familiar faces, to aid the matching task.

## 2.7 Between Experiments Analysis

In order to compare the familiarity advantage across experiments (and across image quality), I next tested how the results of Experiment 3 compared with the results of Experiments 1 and 2. As familiarity was assessed across three levels in Experiment 3, I reanalysed the data from data experiments 1 & 2 in the same way to allow direct comparison.

A 3x3 mixed ANOVA was performed on the results of Experiment 1, 2 and 3. This compared *Familiarity* (3 levels) with *Image Quality* (Experiment, 3 levels [fine, mid & coarse pixelation]). The ANOVA revealed a main effect for *Familiarity*, such that face-matching accuracy increased with increasing familiarity when pooling over *Image Quality*, *Band 1*, *Low* M = 64%, *Band 2, Mid* M = 65%, *Band 3, High* M = 77%, [$F(2, 174) = 35.24$, $p < .001$, $\eta_p^2 = .29$]. Pairwise comparisons revealed that these differences lay between High familiarity and both Mid (mean difference = 11.33, CI = 7.78-14.49, p<.001) and Low familiarity (mean difference = 12.52, CI = 9.81-15.24, p<.001). There was no significant difference between Low and Mid familiarity, p>.05.

A significant main effect was also observed for *Image Quality*. Performance accuracy was highest for the *Fine* images (Experiment 1) (M=78%, SE=1.65, CI=74.78-81.36) then *Mid* pixelation images (Experiment 2) (M=69%, SE=1.65, CI=65.51-72.08), which were both

higher than performance for the *Coarse* pixelation images (Experiment 3) (M=58%, SE=1.65, CI=54.37-60.94), [F(2,87)= 39.37, p<.001, $\eta_p^2$ =.47].

There was an interaction observed between *Familiarity* and *Image Quality* [F(4,174)= 2.94, p<.05, $\eta_p^2$ =.06]. This is due to the graded familiarity effect seen in Experiments 1 and 2 breaking down in Experiment 3 as a result of extreme image degradation.

## 2.8 General Discussion

It is evident from this series of experiments that being more familiar with a face increases a person's ability to tell that face apart from its lookalike, and also to 'tell together' ambient images of the same face. Familiarity aids imposter detection even in the case of poor quality images, with performance increasing in a graded manner when both *Fine* pixelation 'standard' images were viewed (Experiment 1) and when *Mid* pixelation images (30x34 pixels) were used (Experiment 2). However the findings of Experiment 3 illustrate that the graded familiarity starts to break down as the pixelation becomes more *Coarse* (20x30 pixels). Not only does the graded pattern falter in the overall accuracy data, but also the familiarity advantage is lost for different identity trials.

My findings underscore those of previous research and extend them in several ways. I created a challenging face matching task that addressed face-matching performance for ambient celebrity face images and very similar celebrity and lookalike faces. Performance accuracy on this task was even poorer than had previously been established in matching tasks which used cooperative stimuli. In my lookalike task unfamiliar viewers performed with 72% accuracy, which was a level of performance much poorer than in the GFMT (M = 90% long version, 81% short version) which is a standardised face-matching test consisting of cooperative stimuli (Burton et al. 2010). Accuracy dropped even more with degraded image quality. These results reflect that human face-matching performance is

even worse for challenging images than previously established in standardised cooperative matching tests.

My research also adds to the existing literature on familiarity as a graded concept (Clutterbuck & Johnston, 2002, 2004, 2005). Clutterbuck & Johnston divided familiarity into just three bands (high, medium and low) with faces categorized into these bands according to ratings from an independent rater group. The current Experiments 1 & 2 provide more detailed insight into familiarity effects by tracking performance over five levels of familiarity instead of 3. Importantly, in the present experiments, familiarity ratings were based on participants' own rankings of their level of familiarity with a face, unlike in previous studies which have assumed equal familiarity with the familiar face stimuli for all participants. I also tracked performance across changes in image quality, and found first that familiarity does help improve performance even for coarsely pixelated images, but there is a limit to the familiarity advantage, especially its graded nature. These findings fit with previous demonstrations that face recognition and face-matching performance decline as image resolution decreases; with performance eventually falling to chance (Harmon & Julesz, 1973; Bindemann et al. 2013).

The experiments add to the existing knowledge of the familiarity advantage for accurate face recognition. My study provides the first experimental investigation into the familiarity advantage for distinguishing between true match and lookalike faces, finding that familiarity does indeed aid this challenging task. I have shown that the graded familiarity advantage extends to a more detailed breakdown of familiarity levels than had been previously explored. This more detailed analysis was made possible by acknowledging the idiosyncratic nature of familiarity – different viewers know different faces to different degrees – and by allowing this insight to inform the design and analysis.

Too much pixelation destroys the graded familiarity advantage. Although an overall familiarity advantage did carry through to the case of coarsely pixelated images (20x30),

this advantage was carried solely by the same identity pairs. This eventual breakdown notwithstanding, the general robustness of the familiarity effect against declining image quality suggests that familiar viewers are using information other than purely fine details and precise configural information to support their make face-matching decisions.

Celebrities were used as the target faces in the experiments. It is therefore possible that the results of the study would differ at the extremes of familiarity if photographs were taken from people's everyday encounters, rather than celebrities, to provide the target faces. For example, performance may be better preserved if the target face was a family member, or worse if the target face was someone who had never been seen before (participants may have had some prior exposure to the celebrity faces even if they were not aware of it).

As well as their theoretical interest, these findings may also have practical implications. Forensic investigations regularly rely on face images as a means of evidence (Loftus & Doyle, 1992). I found that the more familiar a viewer was with the target face, the better was their ability to reject similar faces. It seems logical that in situations involving identity fraud or poor quality images that a viewer who is of the highest available level of familiarity with the target face would be best placed to judge the identity of the person concerned – and that a little familiarity may be better than none.

In summary, increasing familiarity with a target face increases a viewer's ability to integrate different images of the same person and to distinguish images of different people – even in the context of very similar faces, and poor quality ambient images. I approach a limit to this familiarity advantage, where increased familiarity cannot fully compensate for reduced image quality. In the next chapter, I consider how performance on this difficult face-matching task might be improved.

There is also practical relevance to this experimental series. Performance was identified to be significantly poorer as a result of image quality degradation, yet there are situations when poor quality images are all that are available to aid an investigation. It is therefore of interest to find ways of improving performance for the reduced quality images used in these experiments. This will be investigated in Chapter 3.

# Chapter 3 – Improving Performance

## 3.1 Chapter Summary

In this chapter I test several ways of improving the poor performance for pixelated images seen in Chapter 2. Pixelated images of faces are often encountered in forensic investigations when zooming in on digital images. Thus improving performance in this has highly applied relevance. I show that blurring the pixelated images, and applying crowd analysis can both improve performance on a pixelated matching task. Moreover, performance benefits due to these were additive meaning that both could be used together for even greater performance improvement. Finally, I found that super-recognisers outperformed control participants, even in the extremely challenging task of imposter detection for poor quality images. This is the first time that super-recognisers' performance has been shown to extend beyond good quality images of cooperative stimuli.

## 3.2 Introduction

In Chapter 2, I showed that familiarity can help when dealing with pixelated images, but this familiarity advantage was pushed to its limit when dealing with the coarsely pixelated images in Experiment 3 (20x30 pixels). In forensic situations, the problem of identifying pixelated faces is often encountered because this is the resulting image type from zooming in on CCTV footage (Bindemann et al., 2013) or other digital images (Jenkins & Kerr, 2013). In many such cases, finding a viewer who is familiar with the faces concerned is not a viable option. For example, passport security officers or club bouncers may be required to make identity judgments concerning many individuals in a very short period of time. With poor image quality being a very real problem in applied environments, it would be useful to find ways to improve human face recognition performance for pixelated images across the familiarity continuum. In this chapter I will test several very different methods for improving pixelated face recognition, specifically image manipulations, data analysis and specialist viewer groups.

Most previous attempts to improve face recognition performance have revolved around training, usually using good quality images. However, training approaches have met with little success. In an early example, Woodhead, Baddeley & Simmonds (1979) evaluated a longstanding three-day training course, which aimed to improve face recognition performance of attendees. The course was deemed intensive, consisting of lectures, demonstrations, discussion and practical work. Specific focus was placed on learning isolated features, as the course founders believed this to be the key to successful face recognition. Before this study was conducted, the success of the programme had not been measured. To measure its effectiveness, attendees were tested on face-matching and face *memory* tasks both before and after the three days of training. Their performance on these tasks was compared with a control group who did not take part in the training course. It was concluded that undergoing training did not significantly improve performance in any of the test tasks, with both trainees and controls showing similar mean hit rates between .6 and .9 depending on the task type. In one test training actually led to a significant decrease in performance compared to controls. The authors explained this by suggesting that attention to isolated features may impair face recognition performance rather than improve performance in some instances (Woodhead et al. 1979). This explanation for poorer performance after training is built on the research of Winograd (1976) who found that when participants focused their attention on one specific facial feature, their memory for the face was impaired.

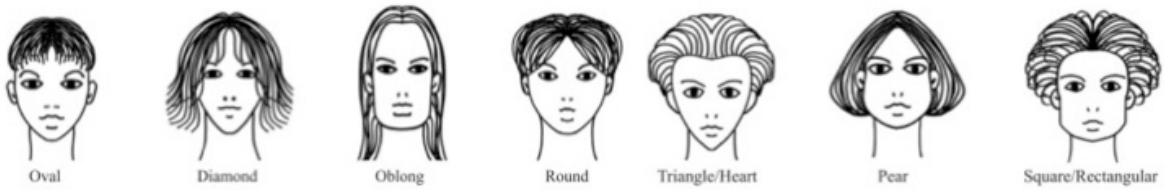


**Figure 3.1** Face shape classification examples provided by Towler et al. (2014).

Overall Woodhead et al. (1979) showed that attendance on a training course which focused on attention to isolated features did not improve face recognition accuracy.

However, it is possible that training in other methods could help. Towler, White & Kemp (2014) assessed the face shape training strategy, which focused on classifying the shape of a face (e.g. oval, round, square) as the basis for successful face recognition (see Figure 3.1). Unfortunately this technique was also not successful in improving face recognition. Participants of the course were no better at the GFMT after undergoing the face shape training than they were before training. The face shape training strategy seems to be fundamentally flawed, as a face does not take on a consistent shape and also, it is difficult to classify faces by shape as there are not particularly clear distinctions between each category e.g. what one person considers a round face another may consider to be oval. This was found to be the case in the study, as different images of the same individual's face were frequently described as having different face shapes. During training participants viewed a series of five different same identity photographs. Each identity was judged as having the same face shape across all five photographs in only 7% of cases. It was noted that the perceived shape of a face was not a diagnostic characteristic of identity, hence explaining why face shape training does not improve face recognition. I return to the issue of face shape in Chapter 4.

Some training studies have shifted away from identifying specific aspects of a face, to more general strategies. White, Kemp, Jenkins & Burton (2014) showed slight but significant improvements in face-matching performance when participants received immediate feedback on their decisions. Here the authors were not concerned with how the viewers made their decisions, but instead focused on notifying viewers on whether identity judgments they made were right or wrong. This worked in the form of participants' receiving a correct or incorrect statement, immediately following their answer submission for each trial of a face-matching task that they completed. The images remained on the screen while the feedback was presented and improvement generalised to new faces shown in the task, but it is not known how long lasting the benefit from the feedback would be.

Taking a rather different approach, Dowsett & Burton (2014) also report some positive effects of working in pairs when making matching identity judgments. Individuals were tested for face-matching performance, and then tested as part of a pair, and finally tested individually again. In the pair judgment condition, the identity decision was made collaboratively after discussion. Working in a pair, lifted performance to the level of the higher individually performing member. Interestingly, the effect of pair working carried over to improve the performance, especially of the weaker pair member, at the later individual testing. This study, along with the work of White et al. (2014), provide evidence for feedback as an important self-regulator of performance accuracy, which may improve face recognition in some situations. Although performance improvements were statistically significant, they were numerically small in both studies.

Although training courses have been found not to improve face recognition performance, and recent lab based studies showed only small improvements, it remains a widely held belief that those who hold jobs that rely on the ability to accurately identify faces, will be better at face recognition than people whose jobs do not rely on this ability.

In reality, it seems that highly trained officials are no better than untrained and inexperienced others at matching faces. Burton et al. (1999) reported that police officers performed at the same level of accuracy as undergraduate students in a task that involved matching poor quality CCTV footage images to comparison face images. More recently, passport officers have been tested on the GFMT. Passport officers' performed with 79.2% accuracy on the task, whereas mean performance for the general population control group was 81.3% (Figure 3.2). There was no significant difference between these performance scores. It could be argued that the GFMT task does not mimic the identity matching task that passport officers perform, which involves comparison of a photograph to a physically present face rather than to another static image. White et al. (2014a) addressed this by testing passport officers on a task that directly mimicked the passport control scenario of matching an image to a physically present face, and also on image-to-image matching performance. Passport officers again performed no better than a control

group of undergraduate students. Relating back to the work discussed above on training, no relationship was found between number of years on the job and performance accuracy. These findings make sense with reference to previous research – if training holds little benefit elsewhere, there is no reason, other than perhaps increased motivation, why similar training would have improved performance for those people whose jobs rely on high face recognition accuracy when they have not helped before.

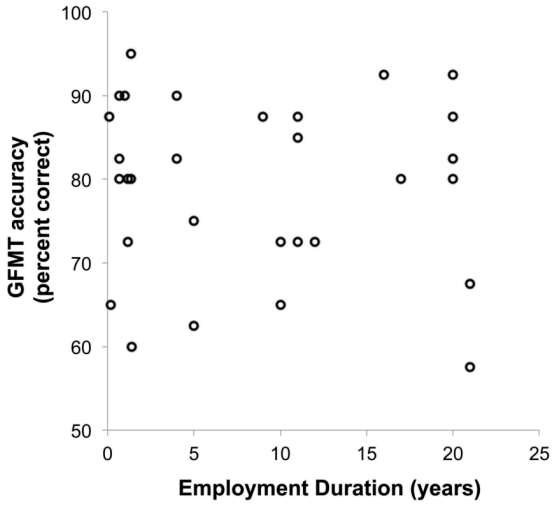


**Figure 3.2** Graph from White et al. (2014) passport officer paper showing the officers' performance accuracy on the GFMT alongside their employment duration. Some police officers performed very highly on the GFMT, these high scores can be found at both ends of the employment duration axis.

Taken together, the previous research on improving face recognition performance suggests that training is unlikely to improve face recognition performance on challenging pixelated images. In this chapter I will attempt to improve face recognition performance using three distinct approaches, none of which rely on training. There are three methods that I will test for improving face recognition performance. First I will investigate whether image manipulations, in the form of blurring pixelated images, can improve performance. Second I will examine whether crowd effects, which is a data analysis technique, can be applied to data that has already been collected to improve performance. Third, I will test whether super-recognisers can be relied upon to make more accurate identity judgments than controls for challenging images. Each of the approaches I use, and past research involving these techniques will be described and discussed in detail within the body of this chapter.

## 3.3 Experiment 4: Blurring Pixelated Images

It was evident in Chapter 2 that although an advantage of extreme familiarity survived for face recognition performance when dealing with coarsely pixelated image pairs, overall performance was poor compared to the prior better quality image pair experiments. Performance was at around chance level for faces that were unfamiliar, and poor for the extremely familiar faces relative to performance for this familiarity level in Experiments 1 and 2 of Chapter 2.

It has been noted by past researchers that ability to recognise a pixelated face can be improved by blurring the image. It may sound somewhat counterintuitive to blur a pixelated image, because blurring removes information from the image. Harmon & Julesz (1973) explain that when an image is pixelated, each square is a result of the average density of the pixels that makes up this area in the original image. There is more of a difference in amplitude between two adjacent pixels than there may have been between two of the pixels in the original image. This difference in amplitude introduces high frequency noise at the pixel edges, making it difficult to extract useful information. Configural information becomes less precise and featural information less detailed. However, when a pixelated image is blurred, high frequency noise is removed and identity is easier to recover (Harmon & Julesz, 1973; Morrone, et al. 1983). Blurring is essentially the same as low pass filtering, in that both processes filter out high spatial frequency information.

Here I examine whether blurring the pixelated images used in Chapter 2, Experiment 3 will improve performance despite the extreme similarity of the face images in each pair. If blurring enhances performance on the task then this technique could be used in applied settings to aid facial identifications in forensic investigations.

I predict that removing high frequency image noise through blurring the pixelated images will improve accuracy as compared against the results of Experiment 3.

## Method

### Participants

A group of 30 undergraduate students (M= 6, mean age = 19.7) at the University of York (who had not taken part in any of the previous lookalike tasks involved in this series of experiments) volunteered as participants for this study.

### Design and Stimuli

As in the previous experiments, this experiment adopted a within subjects design. The variable familiarity was examined at three levels in order to keep consistency with the design of Experiment 3. This allowed me to perform a between experiments comparison of performance for the different image types (pixelated and blurred), comparing the results of this experiment from those from Experiment 3.

### Face-matching Task

The stimuli for this experiment were modified versions of the face images from Experiment 3. To create the stimuli versions necessary for this new experiment, I took the pixelated images used as the stimuli in Experiment 3 and applied a blurring technique to these images using Adobe Photoshop (CS6). The 20x30 pixel face images were blurred at a radius of 3.8 pixels using Photoshop's Gaussian blur function (see Figure 15 for a side by side comparison of the coarsely pixelated and blurred-pixelated stimuli). This blurring level was determined via pilot testing in which two raters assessed changing pixelation levels on a sliding scale and decided by eye on a level that they believed made the image easier to identify.

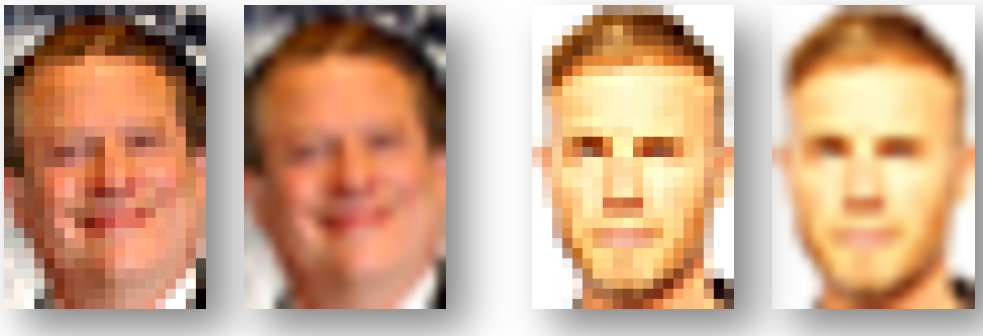

**Figure 3.3** Identical images of Al Gore (left) and Gary Barlow (right) shown as they were presented in each experimental condition. The image on the left for each identity, shows the coarsely pixelated image as presented in Experiment 3. The images on the right, show the image on the left of it, after undergoing blurring, and as presented in Experiment 4.

## Procedure

The procedure was the same as for Experiment 3 except that the images viewed were a blurred version of the coarsely pixelated images. As in the prior experiments, participants ranked each of the celebrity faces in order of familiarity using the familiarity scale (see back to procedure section in Experiment 1 for a detailed description of the face-matching task and familiarity rating scale used for all experiments in this series).

## Results

There was a significant main effect of *Familiarity* on face-matching performance for the blurred version of the task [$F(2,58) = 6.07$, $p<.01$, $\eta_p^2 = .17$]. Participants performed with lowest accuracy for familiarity band 1 (M = 62%, SE = 2.15), with accuracy increasing as familiarity increased for band 2 (M = 65.5%, SE = 2.5) and band 3 (M = 72.3%, SE = 1.81). Performance for familiarity band 3 was significantly greater than for band 1 and band 2, no other differences were significant.

In Experiment 3, the familiarity advantage was found only for same identity trials. In order to check whether blurring recovered the graded familiarity advantage for each type, performance accuracy for the blurred pixelated faces was analysed according to same and different trials using a mixed ANOVA.

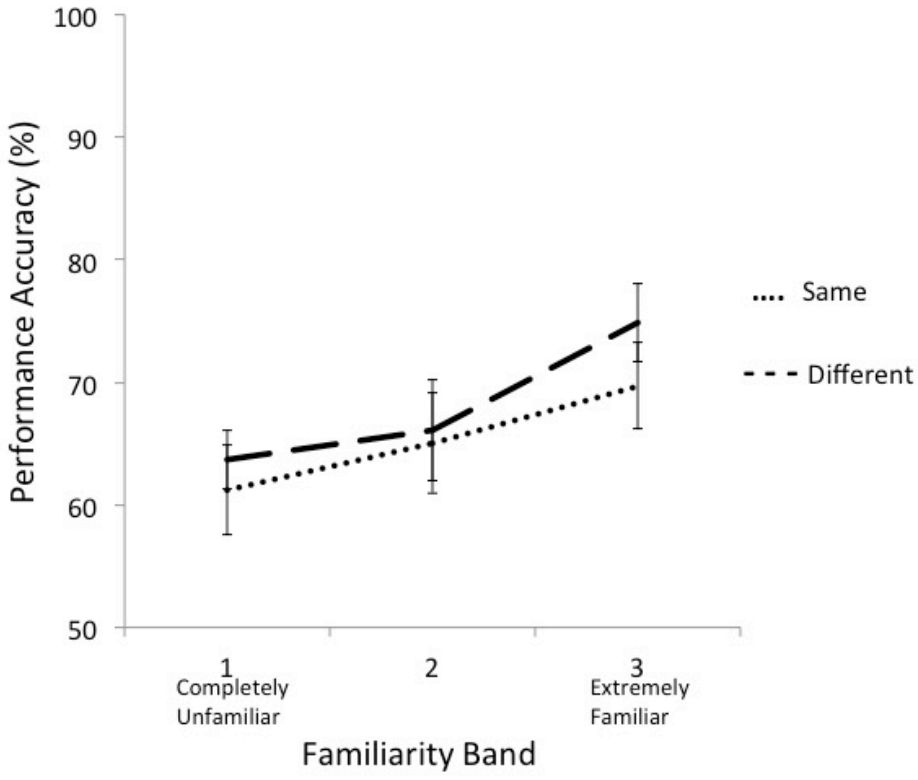


**Figure 3.4** Graph showing performance accuracy on the blurred pixelated task split by same and different person trials. Error bars show the standard error of the mean.

For the *Same* and *Different* identity trial breakdown there was a significant main effect of *Familiarity* [F(2, 58) = 6.02, p<.01, $\eta_p^2$ = .17] but no main effect of *Trial Type* [F(1,29) = .96, p=.33, $\eta_p^2$ = .03] and no interaction between *Familiarity* and *Trial Type* [F(2,58) = .13, p = .87, $\eta_p^2$ = .01]. This shows that for the blurred version of the faces, familiarity improved performance on same person trials and on different person trials (see Figure 3.4).

**Between Experiment Analysis**

To find out whether performance was better for pixelated and blurred faces than for pixelated faces, a between experiments analysis was conducted using a 2 way mixed ANOVA. Factors for the ANOVA were familiarity (within subject factor) and image type, pixelated or pixelated and blurred (between subjects).

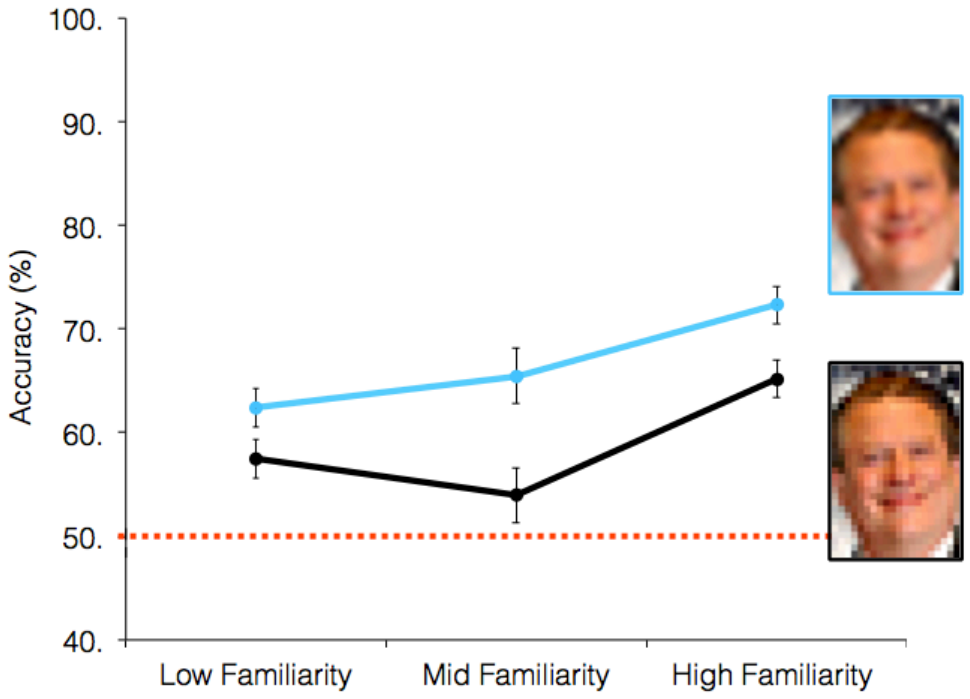


**Figure 3.5** Percentage of correct responses in face-matching task for each familiarity band (low familiarity, mid familiarity, high familiarity), for Experiments 3 (black line) & 4 (blue line). Error bars show standard error of the mean.

Overall, performance accuracy was significantly higher when pixelated images were blurred (M = 66, SE = 1.36, CI = 64. -70) than performance had been in the Experiment 3 which consisted of the coarsely pixelated images (M = 59, SE = 1.36, CI = 56 -62), [$F(1,58) = 16.75$, $p<.001$, $\eta_p^2 = .22$], see Figure 3.4.

Additionally, a main effect of Familiarity was evident from the results [F(2,116) = 12.53, p>.001, $\eta_p^2$ = .18]. Participants were better at recognizing the faces they were most familiar with (M = 69, SE = 1.27, CI = 66-71) than the faces that were of *Mid* (M = 60, SE = 1.89, CI = 56-64) or *Low* familiarity (M = 59, SE = 1.34, CI = 57-62) to them. No interaction was observed between *Familiarity* and *Experiment* [F(2,58)=1.33, p=.27, $\eta_p^2$ =.02].

To keep analysis consistent with that of previous experiments, a three way ANOVA was conducted for *Familiarity* (within-subjects, 3 levels – *Low* familiarity, *Mid* familiarity, *High* familiarity), *Trial Type* (within-subjects, 2 levels – same, different) and *Experiment* (between subject, 2 levels – *Experiment 3*, *Experiment 4*). The ANOVA revealed a significant main effect of *Familiarity* [F(2,116) = 11.2, p<.001, $\eta_p^2$ = .16]. As expected from Figure 2.11 and Figure 3.5 there was significant three way interaction between *Familiarity*, *Trial Type* and *Experiment* [F(2,116) = 3.53, p<.05, $\eta_p^2$ = .06] confirming that *Familiarity* affected performance according to *Trial Type* in different ways across the experiments. No other results were significant.

**Discussion**

Blurring the pixelated images had a significant positive effect on performance accuracy. Overall, performance was 12% better when participants viewed blurred versions of the pixelated images (Experiment 4) compared to when participants viewed the pixelated image (Experiment 3).

Previously, blurring a pixelated image had been found to aid identity recognition of a pixelated face (Harmon & Julesz, 1973; Morrone, et al., 1983). My research found that this blurring advantage extended to matching involving similar images. Blurring the image to a greater or lesser degree may have led to more of an improvement, but I was

primarily concerned on whether or not any blurring would improve performance, which it did.

My experiment also leads to theoretical implications, helping to answer whether identity information is carried in high or low spatial frequency information in a face and more specifically, how this information plays out in the identifying similar faces. When a pixelated image is blurred, the high spatial frequency information in an image is removed. It was previously reported that faces could be identified from blurred pixelated images, which suggests that it is the low spatial frequency information which is important for identity (Harmon & Julesz, 1973; Morrone et al. 1983). My task involved very similar images therefore any differences between the identities was subtle. As blurring improved performance for these images, one interpretation of my findings is that some of the subtle differences between identities were held in low spatial frequencies.

The findings of this study are of great practical relevance. Pixelated images are often the image type which police have to deal with. It has been shown by previous studies that performance on unfamiliar face-matching tasks is extremely poor, especially when the images are of poor quality. Experiments 1, 2 and 3 reiterated this, with the extreme and more challenging case of imposter detection. I found that performance was particularly poor in the imposter detection task for the unfamiliar faces, and performance deteriorated as a function of image quality. Blurring the pixelated images is a simple image manipulation that police could use to aid the likelihood of correctly distinguishing and detecting *same* and *similar* faces.

## 3.4 Experiment 5: Crowd Analysis

In some situations data on identity tasks has already been collected before the challenging nature of the task due to pixelation has been addressed through image manipulation. It is possible that more accurate results of identity judgments can be

obtained from the existing data, through reanalysing the data already collected but in a different way. We have seen already that performance is poor, even for familiar viewers, when the images are coarsely pixelated. Blurring the pixelated images, as above, did improve face recognition performance, however accuracy remained lower than it did for good quality images. A technique called *crowd analysis*, may improve performance accuracy, and could even help in the situation when no viewers are extremely familiar with the faces whose identities are in question.

Crowd analysis is based on the idea that a group performs better than an individual (Galton, 1907; Krause et al., 2011; White, Burton, Kemp & Jenkins, 2013). Rather than relying on a single person's decision, the majority vote can be obtained by pooling these individuals' results to find the group average response. The most commonly given answer is taken as the group answer – simple majority rule. Crowd effects are thus created by pulling together the mean result of a group of people. This mean answer is then taken as the new response and compared against the correct answer.

The power of the crowd over an individual has been acclaimed for quite some time; Aristotle addressed the wisdom of the crowds in his book Politics (Aristotle, published 1920). Later, statistician Sir Francis Galton (1907) demonstrated that a crowd could outperform the individual by calculating the median estimated weight judgment of an ox in a guess the weight of an ox competition at a fair. He found that the median score of a crowd of 800 people provided an answer that was within 1% of the true weight of the ox. Since Aristotle and Galton, the concept of wisdom, or power of crowds has been referred to by many different names. In biological terms, this phenomenon is commonly referred to as swarm intelligence, and refers to the superior ability of groups in solving cognitive problems over the individuals who make up that group. The benefit of swarm intelligence has been more recently replicated through experiments involving guessing the number of marbles or the number of sweets in a jar (Krause et al. 2011; King et al. 2012). An interesting finding in these studies was that groups of low performers could outperform individual high performers. Krause et al. 2011 highlight that swarm intelligence is

beneficial in providing a more accurate answer to a sweet in a jar question, but less beneficial when dealing with questions that require expertise of knowledge – this was tested with regards to predicting coin toss odds in relation to the odds of winning the lottery. For the coin toss and lottery question, groups would only outperform experts if the group size was larger than 40 people. Group sizes varied between 2 and 80, with larger group sizes providing more accurate results.

King et al. 2012 also found that swarm intelligence improved accuracy of sweet guessing judgments, but then adapted the experiment to find out whether providing the individual guessers with additional information (the average guess taken by the people before them, the guess of a randomly chosen previous person, the guess of the person before them or the best guess which had been made before theirs) would sway their judgment. With no additional information, the crowd far outperformed the individuals. Knowledge of any of the previous guesses, except the case of the best previous guess, reduced the effect of swarm intelligence. Access to the best guess led to more accurate results than in each of the other conditions at both and individual and small group analysis level. This suggests that crowd analysis does help, in all situations, but the added information of an expert's guess could help improve accuracy further.

White et al. (2013) were the first to have used the wisdom of the crowd theory to address the problem of face identity judgments, and refer to the technique they use as *crowd analysis*. The principle remains the same; the most common result of the group is taken as the answer (majority vote) instead of averaging responses at an individual response. White and colleagues (2013) report that performance accuracy on the GFMT can be improved by analysing the results of the GFMT (Burton et al. 2010) for crowd sizes of 2, 4, 8, 16, 32 and 64 subjects, rather than looking at the mean overall response at an individual level as previously reported (see Figure 2.3). It is important to note that the authors are dealing with the exact data collected from Burton and colleagues (2010) study, where mean performance for individual accuracy was 89.9%, it is only the type of analysis which has changed. Crowd analysis, for groups as small as four people,

outperformed mean individual accuracy. Group size was also important, with larger groups providing more accurate results of identity matching than when individual results were pooled across smaller groups. The largest group size of 64 led to performance accuracy levels of 99.2%. Further still, crowds outperformed the highest performing individual when the data was aggregated over a group size of eight or more (White et al. 2013).

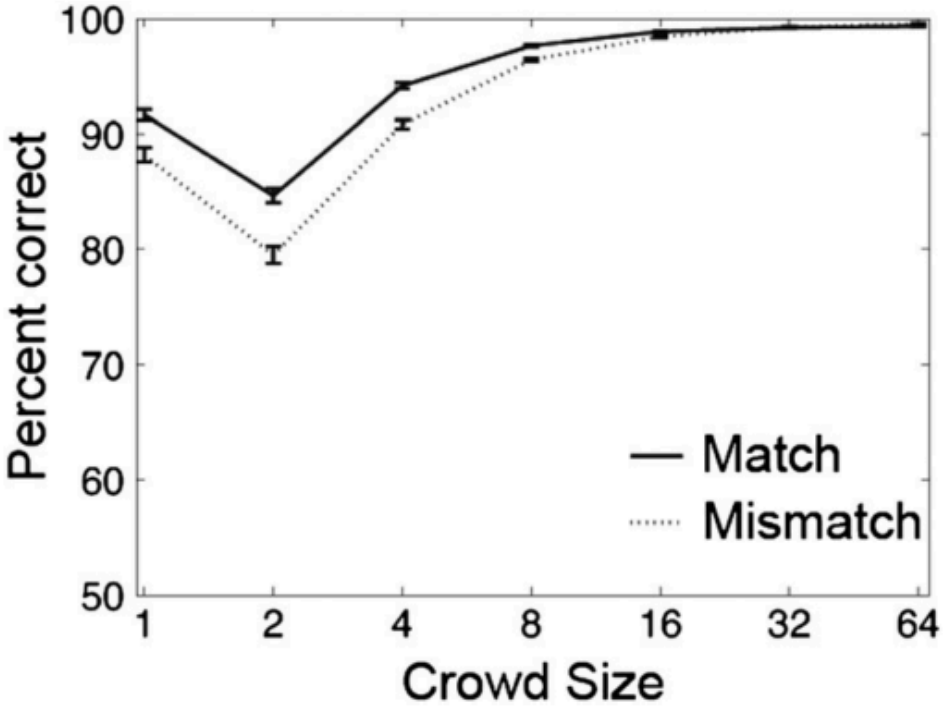


**Figure 3.6** Mean performance on items of the GFMT performance according to different crowd sizes (White et al. 2013). Graph shows performance accuracy broken down by trial type, with results analysed for crowd sizes of 1, 2, 4, 8, 16, 32 and 64.

This study by White et al. (2013) showed for the first time that crowd analysis can improve performance on face identity judgments. The training techniques discussed in the introduction of this chapter focussed on improving performance at the individual level but this led to no or little success. Data analysis techniques may actually allow us to improve the accuracy of identity decisions without the need for training, and even to improve accuracy retrospectively, on tasks for which data has already been collected. This

is useful for new studies also, as data collection methods would not need to change in any way. The crowd analysis technique has proved useful in tasks involving good quality images (e.g. White et al. 2013), but has not yet been applied to data from tasks involving more challenging images.

I will test whether performance can also be improved through crowd analysis for my pixelated lookalike task, which includes very similar and poor quality images. It is not yet known whether the crowd advantage can extend to help in this very challenging case of identity judgment. This is particularly interesting to investigate, as a more challenging task will allow a better understanding of the advantage that crowd analysis can hold. In the GFMT, crowd analysis could only improve individual performance by a maximum of 10%, which brought performance to ceiling level (see Figure 2.3). It is unknown whether crowd analysis could improve performance by more than 10%. Performing crowd analysis on my more challenging task will allow this to be investigated, as the lower baseline performance level in this tasks allows far more room for improvement than in the GFMT.

Based on the success of crowd analysis in previous situations, I predict that performance accuracy on the pixelated lookalike task will improve as crowd size increases.

**Method**

**Data**

The dataset used in this study was the raw data from Experiment 3, the coarsely pixelated lookalike task, (N = 30, M = 8, mean age = 20.2).

**Analysis**

To carry out crowd analysis I calculated the most frequently given response (*same* or *different*) for each item of the coarsely pixelated face-matching task, Experiment 3, across different crowd sizes (i.e. subgroups of participants) and then used this response as the answer for that item. Crowd responses were calculated for all items tested to determine overall percentage accuracy. Results were calculated across crowd sizes of 1, 3, 5 and 15, these crowd sizes were selected as denominators of 30, which allowed the creation of equal group sizes given that there were 30 participants in the study.

**Results & Discussion**

To find out whether mean performance accuracy increased with increasing crowd size, individual performance (crowd size 1) was compared with performance of crowds of 3, 5 and 15 (see Figure 2.4).

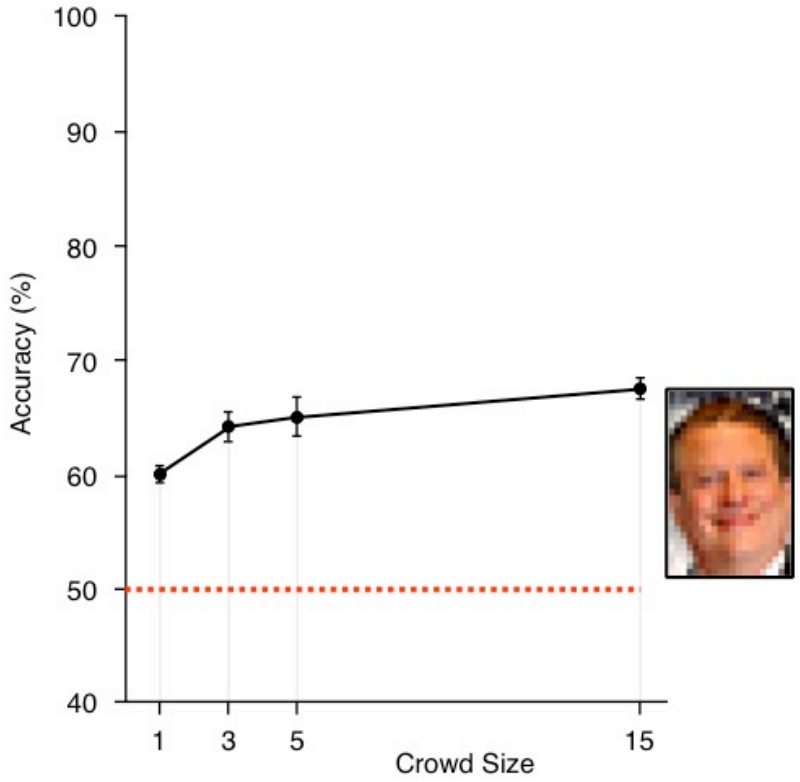


**Figure 3.7** Graph showing the mean accuracy score for crowd sizes of 1, 3, 5 and 15 for the coarsely pixelated lookalike task, Experiment 3. Error bars show the standard error of the mean.

Individuals (M = 60.11 % correct) were outperformed by each of the crowds, with higher performance for larger crowds (crowd size 3, M = 64.16% correct; crowd size 5, M = 65 % correct; crowd size 15, M = 67.5 % correct). The crowds were made up of the same people who contributed to the individual performance analysis, at no point was any individual becoming good at the task - crowd scores were calculated after independent decisions had already been made. These findings show that crowd analysis did improve performance levels, with increasing crowd size leading to better performance.

In keeping with previous analysis, results were broken down into same and different trials. Increasing crowd size increased accuracy for Same identity trials but not for

Different identity trials (See Table 3.1). These results mirror those of trial type reported in Experiment 3.

| | 1(SE) | 3(SE) | 5(SE) | 15(SE) |
|---|---|---|---|---|
| Same | 62.2 (.55) | 66.7(2.85) | 70(3.44) | 73.3(3.33) |
| Different | 58(.59) | 61.7(3.8) | 60(2.11) | 57(5) |

**Table 3.1** Crowd analysis results broken down by trial types (same identity, different identity).

Overall, performance improved through crowd analysis, but did not exceed a 10% increase in performance, which would have been possible given the task difficulty. The coarsely pixelated celebrity lookalike task was a very difficult task, which is shown through the highest group performance being just 70% accuracy. It is possible that performance could be improved even more if I apply the crowd analysis technique to the results from the blurred pixelated faces, Experiment 4. I have shown that blurring improved performance on the lookalike task, possible that crowd analysis on the slightly easier, blurred, version may lead to even better performance than either crowd analysis applied to the coarsely pixelated faces, or blurring the faces, alone.

**Combining Improvement Techniques (Experiment 5b)**

Blurring pixelated images (Experiment 4) and performing crowd analysis to pixelated images (Experiment 5) have both resulted in improvements in performance accuracy compared to performance on the coarsely pixelated celebrity lookalike task (Experiment 3). Both blurring and crowd analysis had been proved successful in improving face-matching accuracy in past studies, yet no work has attempted to combine these techniques, i.e. perform crowd analysis on the blurred pixelated image data.

I predict that crowd analysis will have a similar effect on the blurred pixelated data as it did on the coarsely pixelated data, therefore increasing crowd size for the blurred pixelation data will lead to more accurate results.

## Method

### Data

The dataset used in this study was the raw data from Experiment 4, the coarsely pixelated lookalike task, (N = 30, M= 6, mean age = 19.7).

### Analysis

Crowd analysis was carried out on data from the blurred coarsely pixelated celebrity lookalike task, Experiment 4. The analysis procedure was carried out in the exact same way as in Experiment 4 above.

### Results & Discussion

To assess the effect of crowd analysis on the blurred pixelated images, crowd analysis was calculated and results compared for increasing crowds. Crowd analysis improved upon individual performance and improvements increased as crowd size increased. The crowd size of 1 (M = 67.28% correct) was outperformed by crowds of 3 (M = 72% correct), 5 (M = 75% correct) and 15 (M = 80% correct). These results are illustrated alongside the results of Experiment 5 in Figure 3.8.

Crowd analysis was broken down by trial type for the blurred images (Table 3.2). When the images were blurred, crowd analysis improved performance for both same and different trials.

| | 1(SE) | 3(SE) | 5(SE) | 15(SE) |
|---|---|---|---|---|
| Same | 64.3(.67) | 68(2.29) | 70.56(2.5) | 73.3(3.33) |
| Different | 70.2(.46) | 75(1.87) | 80(2.11) | 87(0) |

**Table 3.2** Crowd analysis results broken down by trial types (same identity, different identity).

Next, results of blurred crowd analysis were compared with the results of the coarsely pixelated experiment crowd analysis.

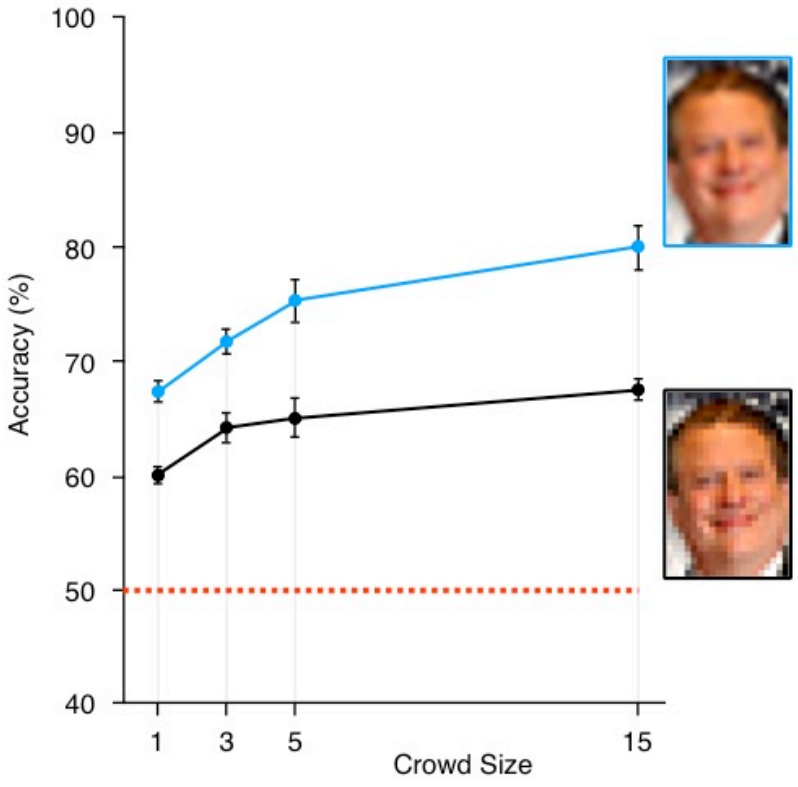


**Figure 3.8** Graph showing the mean accuracy score for crowd sizes of 1, 3, 5 and 15 for blurred version of coarsely pixelated lookalike task (blue line) and the coarsely pixelated lookalike task (black line). Error bars show the standard error of the mean.

Performance gain due to blurring (the difference between the black and blue lines in Figure 2.5) and performance gain due to crowd effects (represented by the slope of the

lines in Figure 2.4) are additive. This pattern suggests that the two gains have independent causes (as demonstrated by the results of Experiments 3 & 4 broken down by Trial Type). It also implies that both techniques can be used in combination to secure the benefits of both. Indeed, using both techniques together leads to a large (20%) accuracy gain in performance, from 60% accuracy for a crowd size of 1 in the coarsely pixelated condition, to 80% accuracy for a crowd size of 15 in the pixelated and blurred condition. The significance of this gain in performance is apparent when compared with the gain of previous techniques such as training, which have improved performance accuracy on previous face-matching tasks by around 5%. Overall, crowd analysis and blurring has taken mean performance of individual viewers from 60%, which was low compared to both unfamiliar viewers on easier (unedited) no pixelation versions of the task, and compared to the performance of familiar viewers on any of the tasks, up to the level of mean performance accuracy level of viewers who were fairly familiar with the faces concerned in the easiest, no pixelated version of the task, Experiment 1. This highlights that blurring and crowd analysis applied in combination greatly improving performance for difficult images.

My results are important with regards to methods for face improving matching accuracy. In past research, large increases in performance have been linked to the use of familiar viewers in a task (e.g. Burton et al. 1999). The success of blurring and crowd analysis techniques does not rely on familiarity with the faces in the matching task. Crowd analysis, combined with blurring where appropriate, could thus aid judgments of very difficult identity decisions and provide a solution to improving face recognition performance that does not rely on familiarity with a face.

## 3.5 Experiment 6: Observer Factors

Unfamiliar face recognition ability is generally poor. We have seen this both in past research (Burton et al. 1999; Bruce et al. 2001; Kemp et al. 1997) and in the results of the thesis up to this point. There is however large variation at an individual level in face

recognition performance – some people consistently perform with very high levels of accuracy, some with very poor levels of accuracy, and others at intermediate levels. This was seen in the study by White and colleagues (2014), which tested face-matching performance for passport officers. It was reported that although training and years of experience made no difference to performance, there was a huge amount of variation in performance amongst the passport officers themselves (White et al. 2014a). Face recognition is therefore generally thought of as an ability that lies on a spectrum. At the extreme lowest end of the face recognition ability spectrum are people with congenital prosopagnosia (Duchaine, 2011). Congenital prosopagnosics have no identified brain deficits, yet experience clinical level of difficulty with recognising faces (Berham & Avidan, 2005; Duchaine & Nakayama, 2006). At the other end of the spectrum are people who consistently perform with exceptionally high accuracy. These people have been termed 'super-recognisers' (Russell, Duchaine & Nakayama, 2009). Super-recognisers may provide a solution to the face recognition problem in practical settings. In theory, super-recognisers could help to eliminate the number of face-matching errors made, in situations where face-matching is of high importance. If super-recognisers truly make fewer errors than others, which is what these past studies have shown (e.g. Russell et al. 2009), then employers should consider testing candidate's face recognition ability, and making super-recogniser performance a requirement for the job roles for which accurate face identification carries high importance.

Members of a highly specialised expert forensic group in the USA have however been found to outperform controls on tasks involving face recognition. The ability of these group members has only recently been identified. Indeed, White et al. (2015b) showed that forensic examiners in the USA performed more accurately in three face-matching tasks than control groups of trained experts in biometric systems (referred to in the study as 'controls') and also undergraduate students (referred to as 'students'). The trained experts in biometric systems were highly motivated with the task, suggesting that differences in performance were not due to differences in motivation. The face-matching tasks used were the GFMT and two new tests, one which was specially designed to be challenging to both computers and humans [EFCT] and the other made of stimuli which

contained cues that would be of use to humans only [PICT]. It is unclear so far whether people gravitate towards the job of forensic examiner as a result of innate face recognition ability, or whether the extremely specialised training helps these forensic officers to perform highly on the face recognition tasks. As highlighted earlier in this chapter, training has a poor track record for improving face identification performance (Woodhead et al., 1979; Towler et al., 2014). The lack of success of previously evaluated training methods would incline one to believe that an innate ability for face recognition is key. It seems from the findings that the forensic examiners were using different techniques to the comparison groups to make their identity judgment. This could be a result of innate ability or training. It would be interesting to find out exactly what the training for these forensic examiners involved, and this knowledge may help to answer whether the examiners' face recognition ability is innate or learnt.

Additionally, further testing and analysis of the data collected from White and colleagues' (2014) study on passport officers' face-matching performance, revealed that a subset of the passport officers were identified as particularly high performers. These highly performing officers were in a specific branch of the passport office, in a job role where their ability to successfully match true match faces and identify false matches was of paramount importance. Thus, high face recognition ability was even more important for their job role than for the roles held by other passport officers tested (White et al. 2015a). It remains unknown whether the officers came to their specific role due to being naturally best suited, or if they learnt the skills required on the job. This has not been investigated directly to date. As years of training was recorded, and found to have no effect on performance accuracy, it may be more likely that innate ability guided these passport officers into their specialised job, where face accurate face recognition is extremely important. As in the case of the US forensic examiners, more information about the training programmes and ideally also new data that specifically tracks only the members of these highly specialised expert groups over the course of training and job experience, could help to answer whether ability has been innate, learnt or perhaps even a combination of both.

The performance of super-recognisers had until recently not been tested and directly compared to controls by standardised tests. Instead, super-recognisers were identified from large groups of people tested on standardised tests, as those with the very highest scores. Bobak, Dowsett & Bate (2016) specifically tested the performance of self identified super-recognisers to find out whether they were better than a student control group at two standardised tests - the GFMT and a face-matching task involving models. The models task is challenging, as the same models can look very different across different images. All of the model faces were unfamiliar to the viewers, therefore this study tested purely unfamiliar face recognition performance of super-recognisers. The authors reported that their group of 7 self-identified super-recognisers were better than a student control group at the two tasks. This finding suggests that people who have exceptional face-recognition ability are in fact aware of this ability, as their superior performance to others was evident when tested on standardised face identification tasks.

**Super-Recognisers within the MET Police**

The evidence provided overwhelmingly supports the use of super-recognisers as a possible solution to the face identity problem. It therefore seems sensible that face identity professionals consider employing super-recognisers. This advice has been taken by The Metropolitan Police who have recently established a super recogniser team within their force. This has been created by internally recruiting officers who have particular interest in face recognition and performed well in their undisclosed face recognition test. These officers regularly try to identify faces from pixelated CCTV photographs. During the establishment of the super-recogniser team I was given the opportunity to test four of these MET police super-recognisers in their normal working environment, to see whether they outperformed our undergraduates on the lookalike task. The lookalike test is particularly relevant for measuring face recognition ability of the police super-recognisers as it tests performance of faces of differing levels of familiarity – from unfamiliar to extremely familiar. The police have to deal with cases from both levels of this spectrum depending on whether the individual involved in the crime is a known repeat offender or committing their first crime. Due to time constraints I could test performance on just one

version of the test. I wanted to test performance on one of the pixelated lookalike tasks, as Chapter 1 demonstrated that even familiar undergraduate viewers do not score with perfect performance on these tasks. Using pixelated images would therefore avoid ceiling effect for matching trials. Pixelated images are also an image type that the police have to deal with as part of their investigations, meaning that the task holds practical relevance. The mid pixelation level (30x45 pixels) was chosen so that predicted performance would be at a level expected to be above chance as based on previous findings (the higher pixelation level reduced performance to chance for all but extremely familiar faces in our experiment on undergraduate students), this also guaranteed that I had an already existing comparison group whose performance was above floor (>50% accuracy).

**Participants**

Participants were 4 super-recognisers (M=4, mean age = 40) from the Metropolitan police force super-recogniser team, New Scotland Yard Central Forensic Image Unit, London. Performance was compared with that of our 30 undergraduate students tested in Experiment 2 (M = 11 male, mean age = 19.7).

**Design & Stimuli**

The stimuli were images of the 30 celebrities and lookalikes for these celebrities used in Experiment 2. As the comparison viewers were generally younger than the super-recognisers, I took care to ensure that any differences in performance could not arise through some celebrities being more familiar to one group or the other.

As in Experiment 2, all images were presented in low resolution (30 pixels wide x 45 pixels high). This resampling also served the purpose of reducing matching images to the level that we expect to avoid ceiling effects in the matching trials. Examples of the stimuli are shown in Figure 3.9.

**Figure 3.9** Example trials from the PLT. Images on the left show different identities (with the imposter face on the right). Images on the right show the same identity.

### Procedure

The procedure was similar to that used in Experiment 2. The police super-recognisers viewed the images in a printed booklet rather than on a computer screen. Super-recognisers were presented with a printed booklet containing 60 trials, half of which showed the same identity and half of which showed different identities. They were asked to make a same/different judgement for each trial. Following the face-matching task participants used a numerical scale to indicate their level of familiarity with each celebrity whose face had been viewed in the task (from 1 [completely unfamiliar] to 10 [extremely familiar]).

Police super-recognisers were also tested on the GFMT and the models face-matching task (MFMT) which were designed and administered by other researchers from the University of York Facelab.

## Results

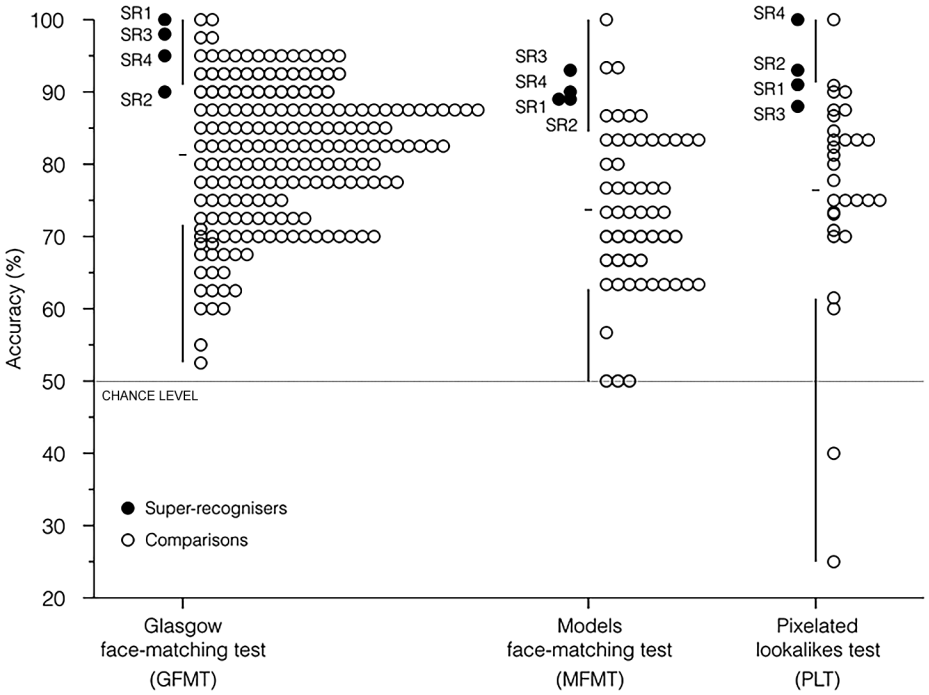


**Figure 3.10** Performance of police super-recognisers and comparison viewers. Performance of super-recognisers (SR1–4; black) and comparison viewers (white) on three different tests of face recognition—the GFMT (left column), the MFMT (middle column), and the PLT (right column). Vertical lines indicate the range of scores for comparison groups, the deleted portion of the line shows the standard deviation, and the horizontal notch shows the mean. In all three tasks, chance performance is 50%.

In Experiment 2, results were broken down according to 5 levels of familiarity, with results analysed according to these familiarity quintiles. Due to the small number of super-recognisers that I was able to test, this time I broke familiarity down into two groups (highly familiar [80-100 on familiarity scale] and less familiar faces [all other ratings]) to ensure sufficient numbers of faces were placed in each familiarity bin.

Analysis was broken down according to the two familiarity bands, with the first tests being performed on the subset of faces that viewers rated as highly familiar. Interestingly, super-recognisers gave high familiarity ratings (80–100) to a very high proportion of faces compared with controls (70% of faces for super-recognisers; 37% of faces for controls).

The police super- recognisers consistently performed with far greater accuracy than student participants. Overall, for highly familiar faces the controls scored just 76% correct, whereas the police super-recognisers scored with 93% accuracy in the lookalike test. All super-recognisers performed much better than the control mean, with one super-recogniser performing perfectly. This performance is shown in Figure 3.10 (for full results of the GFMT and MFMT see Robertson, Noyes, Dowsett, Jenkins & Burton, 2016).

To better understand the super-recogniser advantage, I conducted two further analyses. First, I compared accuracy of super-recognisers and controls on faces that they rated as less familiar (0–7 on the 10-point scale; i.e. those not included in the above analysis). Police super-recognisers outperformed controls on these faces too (76% accuracy for super-recognisers; 66% accuracy for controls), implying that their performance advantage holds across the whole familiarity continuum. Second, I analysed the control participants' data for an association between i) the proportion of faces that were given high familiarity ratings and ii) the level of accuracy on those highly familiar faces. A significant positive correlation was found between these two measures [$r(28) = 0.39$, $p < .05$], such that the highest performing controls were qualitatively similar to the super-recognisers. These analyses support the original comparison, i.e. the super-recognisers are performing at well above the levels of controls, even when group differences in famous face familiarity are taken into account.

Result breakdown by trial type demonstrated that SRs outperformed controls at both same and different identity pairs, for both the low and high familiarity groups (See Table 3.3).

| | % Accuracy Low Familiarity (SE) | % Accuracy High Familiarity (SE) |
|---|---|---|
| SRs | | |
| Overall | 75.72 (8.52) | 93 (5.12) |
| Same | 83.65 (11.78) | 95 (2.84) |
| Different | 67.79 (19.09) | 90 (3.14) |
| Controls | | |
| Overall | 66.4 (1.73) | 76.36 (2.81) |
| Same | 68.98 (2.17) | 81.34 (3.61) |
| Different | 64.10 (2.97) | 70.18 (3.54) |

**Table 3.3** Performance accuracy broken down by viewer group and trial type.

**Crowd Analysis on Super-Recogniser Data**

Earlier in this chapter I demonstrated that crowd analysis improved performance accuracy on the coarsely pixelated face matching task. I also confirmed that combining methods of improvements had additive benefits on performance accuracy. To continue the investigation of the effect combining improvement methods, I performed crowd analysis on the super-recogniser data. This analysis incorporated all items, regardless of familiarity with the item (as in the crowd analysis earlier, Experiment 5). Mean performance accuracy for crowd size of 1 (individual analysis) was 80%. Crowd analysis improved performance, with mean performance accuracy levels of 88% for a crowd size of 4 (see Table 3.5). This performance cannot be directly compared to the crowd analysis in Experiment 5, as the police super-recognisers completed the mid level of pixelation task rather than the coarsely pixelated version. Instead new analysis of Experiment 2 (the mid pixelation level face matching task) was conducted using crowd analysis. Crowd results for the mid pixelation group, Experiment 2, are shown in Table 3.4. Increasing crowd size increased performance accuracy, from 62% for a crowd size of 1, to 83% for a

crowd size of 15. These results highlight the exceptional face recognition ability of the super-recognisers, as a crowd size of just 4 super-recognisers outperformed the largest crowd size of 15 controls (Table 3.5).

| Crowd Size | 1(SE) | 3(SE) | 5(SE) | 15(SE) |
|---|---|---|---|---|
| Same | 66.33 (2.14) | 81.33(3.56) | 83.88 (3.26) | 88.3(5) |
| Different | 58.77(2.1) | 69 (5.24) | 72.2(6.81) | 78.3(1.67) |

**Table 3.4** Crowd analysis results broken down by trial types (same identity, different identity) for the mid pixelation (control) Experiment 2.

| Crowd Size SRs | 1(SE) | 4 |
|---|---|---|
| Same | 86.66 (3.85) | 100 |
| Different | 74.12(7.67) | 76.67 |

**Table 3.5** Crowd analysis results broken down by trial types (same identity, different identity) for the SRs.

**Discussion**

This is the first time that police super-recognisers' performance has been assessed using standardised tests. Superior performance of super-recognisers in the pixelated lookalike task suggests two important conclusions. First, the testing systems used by the police to recruit super-recognisers are indeed successful in selecting people with exceptional face recognition ability. Second, these super-recognisers' ability is not limited to good quality images; super-recognisers are superior to others in face recognition ability even in extremely image conditions involving extreme similarity between target faces and foils. The performance of super-recognisers can be improved even further through crowd analysis. The largest crowd size of 4 super-recognisers was relatively small, but outperformed even the largest crowd size of 30 control participants.

## 3.6 General Discussion

I have shown in this chapter that the poor performance found for the highly challenging pixelated stimuli images can be improved upon through use of a variety of techniques. Performance accuracy was improved by blurring pixelated images, combining results using the crowd analysis technique, and using the results of super-recognisers. Blurring and crowd analysis, and also super-recognisers and crowd analysis can be used in combination for additional benefit.

Each of the techniques that I investigated had shown promise in improving performance in previous tasks of face recognition. However, none of the techniques had been applied to such a challenging identity matching scenario. Blurring had been studied in terms of face recognition rather than identity matching (Harmon & Julesz, 1973). The only study of face recognition to incorporate the crowd analysis technique analysed results from a cooperative face-matching task involving good quality images (the GFMT). Super-recogniser performance has also been studied in terms of good quality images, for unfamiliar faces only (Bobak et al., 2016). Past studies have also looked at each of these techniques exclusively for unfamiliar faces. My findings support the body of existing research, adding that the benefits of all of these techniques extend to very challenging pixelated image conditions regardless of familiarity with a face. I also examined the effect of combining techniques in the combination of blurring combined with crowd analysis, and super-recognisers combined with crowd analysis. As far as I know, these different techniques have never before been applied in combination. Combining the techniques led to even further improvements than using any of the techniques in isolation. This suggests that each of the techniques improve performance in a way that is different from each of the other methods. Blurring reduced high spatial frequency noise from a pixelated image, crowd analysis relies on the performance of the crowd being more accurate than the individual and super-recognisers perform highly, but the methods that super-recognisers use are unknown and may link to innate ability. As each of these methods aid the face

recognition problem in a different way, when methods are combined, results reflect the additive benefit of each.

Many of the strengths of the results of this chapter carry over from the challenging nature of the stimuli set created. To reiterate Chapter 1, the image pairs in the face-matching task consisted of true match (two different images of the same celebrity) and false match (one image of a celebrity and one of a lookalike for that celebrity). The stimuli used in this chapter are pixelated, making a task that addressed a difficult version of a practical problem. This design allowed us to test the improvements to face recognition in a more challenging situation than ever before. In the past, ceiling effects had haltered improvements. The more difficult task that I created allowed the effects of combinations of techniques to be recognised, as base level performance was much poorer than on previous tasks, the improvements made by techniques used alone and in combination could be measured as even combined techniques did not bring performance to ceiling on the pixelated celebrity lookalike face matching task.

I demonstrated that combining techniques helped to improve results, even more than when any one technique was used in isolation. However due to the limited testing opportunity with the super-recognisers, I do not have results for the super-recognisers on the blurred pixelation task, and consequently could not perform crowd analysis on blurred pixelation data for super-recognisers. To more fully understand the effect of combining the three improvement methods I investigated, it would be interesting to test whether performance could be improved even more through all three techniques applied in all combinations i.e. testing the SRs on the blurred version of the task and then in addition to this applying crowd analysis to the results.

Each of these techniques, and the combinations addressed in this chapter could be used in applied scenarios to achieve more accurate face identity decisions. Blurring, crowd analysis and the performance of super-recognisers provide ways of improving performance where past methods of training have failed. These techniques are far easier

and less time consuming to implicate than training, furthermore crowd analysis has the unique advantage of being applicable to improve performance levels using pre-existing data. These methods improve face-matching performance without relying on familiarity with the faces that are available for comparison, making them particularly useful in applied situations where familiarity with the faces concerned is not always a viable option.

Up until now, this thesis has focused on face recognition for challenging stimuli based on incidental differences in appearance across multiple images of a face and incidental similarity in appearance between different person trial images. This has shown that whilst it is possible to incidentally look like somebody else, it is also possible for images of the same person to incidentally look like different identities across images (e.g. Jenkins et al. 2011). All images investigated so far have been ambient in nature, however there could be particular images changes which cause these situations to occur. In particular, the distance from which an image is taken from may influence the appearance of a face, and in turn influence the perceived identity. Chapter 4 will explore this effect of changing camera-to-subject distance, looking at both the effect on the configural information in a face as portrayed in an image, and whether any changes of configuration effect face matching ability for unfamiliar and for familiar viewers.

# Chapter 4 – Changing Camera-to-Subject Distance

## 4.1 Chapter Summary

The experiments in this chapter investigated effects of camera-to-subject distance on configural properties of the face image. Changes in camera-to-subject distance produce non-linear changes in face measurements across images. These changes reduced accuracy in an unfamiliar face-matching task by making same identity images look less like themselves. Identity matching performance was much poorer when unfamiliar viewers compared photographs taken from differing distances, than when the comparison images were taken at the same distance. Familiar viewers were far less affected by this distance change and performed at very high accuracy levels in both conditions. Distance cues compensate camera-to-subject change, suggesting the operation of perceptual constancy mechanisms in the high level visual domain of face shape.

## 4.2 Introduction

The earlier chapters of this thesis, in line with previous research, have shown that face-matching performance is generally poor for unfamiliar viewers. Challenging images in the foregoing experiments (Chapters 2 & 3) resulted in even poorer performance than previous experiments based on cooperative stimuli. So far in this thesis I have focussed on challenging image performance for ambient face images. These images help to capture natural variability in the appearance of any given face. For example, facial expression and pose may change across images, and the environment that face photographs are taken in, can differ. The effects of such factors have previously been examined in isolation, and it is well reported that these superficial image changes can result in impaired performance on tasks involving identity judgment (Bruce, 1982, 1994; Johnston, Hill, & Carman, 1992; Troje & Bulthoff, 1996; O'Toole, Edelman & Bulthoff, 1998; Bruce et al. 1999) However, one interesting and potentially important factor that has received little attention is the effect of camera-to-subject distance on identity judgments.

A few isolated findings suggest that changes in camera-to-subject distance may well affect identification accuracy. In an interesting paper on optics, Harper & Latto (2001) photographed five models (two male, three female) from five different distances (0.32m, 0.45m, 0.71m, 1.32m & 2.70m) and standardised the size of the resulting images (see Figure 4.1). Face images taken further away were visibly flatter, giving the models a heavier appearance and the implication is that the faces appeared to have different shapes in these different distance conditions. Participants gave higher estimates of the models' weight as camera-to-subject distance increased. In a later study (Bryan, Perona & Adolphs, 2012) ratings of trustworthiness, competence, and attractiveness were lower when the camera-to-subject distance was reduced (specifically when the photographs are taken within personal space). Taken together, these findings suggest that facial appearance changes as a result of camera-to-subject distance, yet no studies have looked at how these changes affect performance in tasks involving identity matching judgments. This omission is perhaps surprising, given the emphasis on configuration in the face recognition literature. The configural processing account holds that each face has a unique configuration that is learnt, and that it is knowledge of this configuration that allows us to tell familiar faces apart. This position seems to require that the configuration of a particular face stays constant across images, but the findings of Harper & Latto (2001) and Bryan et al. (2012) suggest that it does not. Given that camera-to-subject distance is rarely kept constant in practical applications of identity matching, effects of camera-to-subject distance on face identification would also be interesting from an applied perspective.

**Figure 4.1** Changes in face shape resulting in differing weight judgments as photographs were taken from far distance (left) to near distance (right), example taken from Harper & Latto (2001).

In a compelling illustration of this Burton et al. (2015) assessed the stability of facial configurations by measuring distances between features in multiple images of three famous politicians (David Cameron, Barack Obama and Angela Merkel), captured from unknown, but presumably different distances (see Figure 4.2 for example). They found that distances between features varied as much between photos of the same person as between photos of different people. This observation seems to challenge a straightforward configural processing account of face recognition and face learning. However as the camera-to-subject distance for these examples was unknown, a more formal assessment of its effects was not possible. Moreover, as camera-to-subject distance was not the exclusive focus of that study, variation in other factors such as pose could also have affected the configural measurements.

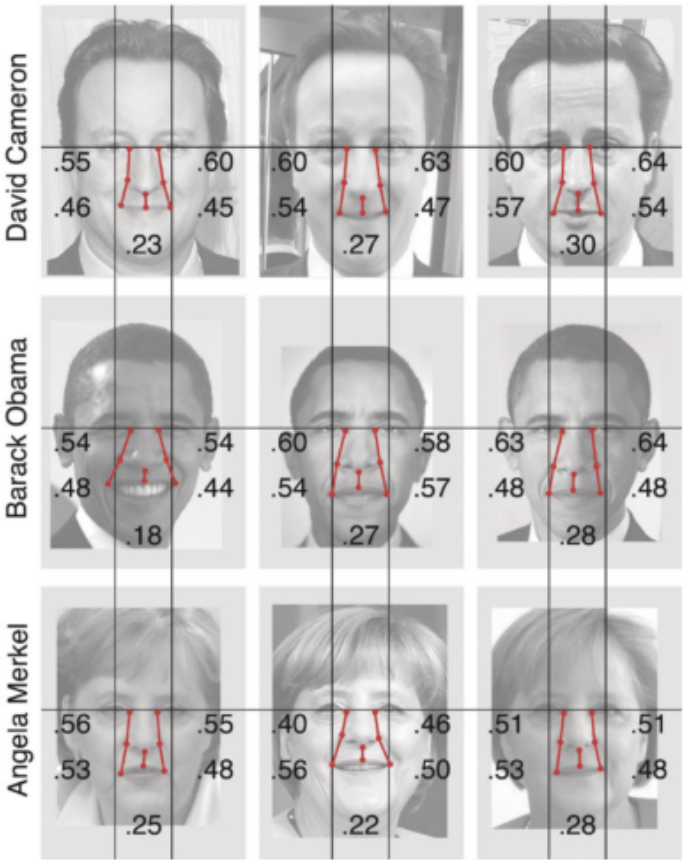


**Figure 4.2** Example of measurement figure taken from Burton et al. (2015). Images are standardised so that interocular distance is the same. Metric distances are expressed as proportions of standardised interocular distance.

It is important to note that the politicians were easily identifiable to familiar viewers in any of the face images, despite the fact that these images varied in configuration. It is a well established finding that unfamiliar face recognition can be derailed by superficial image changes, whereas familiar face recognition is robust against such changes (Burton et al. 1999; Jenkins et al. 2011). This contrast has been linked to familiar viewers having more perceptual experience with the range of appearances that the face can take (Jenkins et al. 2011). Assuming that familiarity interacts with distance-related changes in a similar fashion, it seems likely that these too will impair performance more for unfamiliar viewers than familiar viewers.

Although no previous studies have examined effects of camera-to-subject distance in an identity matching task, a study by Liu (2003) investigated the effect of camera-to-subject distance on face recognition accuracy in a memory task. A face image of each identity used as stimuli in this experiment was edited using Matlab software to create two versions of each face image – one which reflected the appearance that the face would take at a far distance, and one that reflected the appearance of the face from a near distance. Participants viewed face images in a learning phase, and at test were shown either the identical image, or the image altered to reflect the distance which differed from the test image. Liu (2003) showed that the same image of a face is harder to recognise as having been seen before if the image is digitally altered to reflect the changes that would result from altering camera-to-subject distance. My question of interest is different to this in at least two important ways. First, manipulating a single image of a face is not the same as presenting two different images of a face (Bruce, 1982; Jenkins et al. 2011). There is both behavioural (Bruce, 1982) and neural evidence for this (Bindemann, Burton, Leuthold and Schweinberger, 2008). Bindemann et al. (2008) reported that the brain responses from the N250r (an event related brain response generated from the fusiform gyrus) differ depending on whether multiple images of the same face are presented, or if repetitions of the same face image are viewed. Perhaps more importantly with reference to Liu's (2003) study, the N250r response was the same for repetitions of the same face image and for when a digitally altered version of this

image (stretched) was presented. In addition to this, fMRI adaptation has shown that neural representations in the Fusiform Face Area (FFA) to images of the same face does not vary with changes in size, expression or pose of the face in the image (Grill-Spector et al. 1999; Andrews & Ewbank, 2004). These results highlight that responses to stimuli may differ when simply altering the same image of a face digitally, compared to presenting multiple images. Liu's stimuli really created an image recognition task, as images were changed only by simulated distance manipulation.

Second, I am interested in perceptual matching, as distinct from recognition memory. Perceptual matching is interesting in its own right because it allows us to set aside aspects of task difficulty that arise from the fallibility of memory, and to focus on those difficulties that remain at the perceptual level. Face matching tasks also model the task faced by forensic and security officials of determining identity from multiple face images.

I will conduct 3 studies to investigate the effects of changing camera-to-subject distance on facial appearance in an image, and the consequent effects of those image changes on tasks involving identity judgment. I will first characterise configural changes across multiple images of the same individuals taken at known distances. I will then investigate whether these changes translate into difficulties in face-matching accuracy for familiar and unfamiliar viewers. Finally I will evaluate whether the visual system compensates for distance related changes in the face image when distance cues are available.

## 4.3 Experiment 7: Facial Configuration Measurements

The purpose of this study was to relate camera-to-subject distance to facial configuration. The apparent size of an object changes with viewing distance, in the sense that the size of the retinal image changes. Linear changes in the size of a face image (e.g. rescaling a

photograph) do not affect configural information because they do not alter the relative distances between features. Consistent with this observation, both behavioural and neuroimaging studies have found that face recognition is unaffected by linear rescaling (Grill-Spector et al. 1999; Andrews & Ewbank, 2004). For 3D objects (e.g. live faces as opposed to face photographs), the optical situation is different. Changes in camera-to-subject distance generate non-linear changes in the image, such that different parts of the image are affected to differing degrees (Pirenne, 1970). For convex objects, including faces, distant viewing leads to flatter appearance, whereas closer viewing leads to more convex appearance (see Figure 4.1). To relate this transformation to the notion of configuration in the face perception literature, I measured distances between key facial features in photos that were taken at different viewing distances. My expectation was that, as a reflection of the flat-to-convex variation, the change in viewing distance would affect measures near the edge of the face more strongly than it affects measures near the centre of the face.

**Photographic Procedure**

The images used for all of these studies were face photographs of 18 final year undergraduates at the University of York. To allow construction of face-matching experiments (Megreya & Burton, 2008; Burton et al. 2010) these models were photographed in 2 separate sessions, one week apart. In each session, each model was photographed at 2 distances—*Near* (camera-to-subject distance = .32m) and *Far* (camera-to-subject distance = 2.7m), following Harper & Latto (2001). This resulted in 4 photographs for each of the 18 models: Week 1 Near, Week 1 Far, Week 2 Near, and Week 2 Far (72 photos in total). All models were photographed with a neutral expression using an Apple iPhone 5 on default settings. Photos were then cropped around the head to remove clothing and background. For anthropometric analysis, all images were resized to an interocular distance of 150 pixels, preserving aspect ratio (see Figure 4.3).

**Anthropometric Analysis**

For each model in each condition, I measured 5 feature-to-feature distances that have been specified in the configural processing literature: corner of eye to edge of nose (left and right; Leder & Carbon, 2006), corner of nose to corner of mouth (left and right; Leder & Bruce, 2000), and nose to mouth (Leder & Carbon, 2006). Precise anatomical definitions were as used by Burton et al. (2015). The corner of the eye is defined as the centre of the canthus, the corner of the nose as the lateral extent of the nasal flange, and the corner of the mouth as the lateral extent of the vermillion zone. Figure 4.3 shows these measurements for one model. All distances are expressed in units of standardized interocular distance.

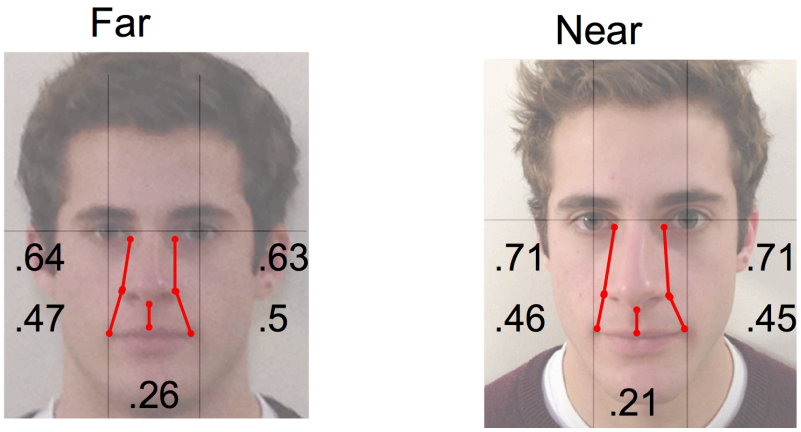


**Figure 4.3** Example of the measurements taken for two of the photos of one model. Measurements taken were the distances between: left eye to nose, right eye to nose, left nose to mouth corner, right nose to mouth corner and centre of nose to mouth.

**Results and Discussion**

For each of the 5 feature-to-feature metrics, I performed a separate 2x2 ANOVA with the within-subjects factors of Photographic Session (*Week*) (Week 1 versus Week 2) and Camera-to-Subject Distance (*Distance*) (Near versus Far). Results of these analyses are summarized in Table 4.1.

| | Week | | | | | | Dist | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg1 | Avg2 | Diff | F | p | ES | AvgN | AvgF | Diff | F | p | ES |
| EN(L) | .57 | .57 | .00 | .05 | >.05 | .00 | .58 | .56 | -.02 | 1.67 | >.05 | .09 |
| EN(R) | .58 | .57 | .01 | .20 | >.05 | .01 | .59 | .57 | -.02 | 3.53 | >.05 | .17 |
| NM(L) | .42 | .41 | -.01 | 1.58 | >.05 | .09 | .38 | .45 | .07 | 39.82 | <.001 | .70 |
| NM(R) | .42 | .41 | -.01 | 1.23 | >.05 | .07 | .38 | .45 | .07 | 42.97 | <.001 | .72 |
| NM(C) | .22 | .23 | .01 | 1.10 | >.05 | .06 | .20 | .24 | .04 | 17.26 | <.001 | .5 |

**Table 4.1** Table showing mean measurements for each photograph condition. EN stands for ear to nose measurement, and NM represents nose to mouth. The letters following denote the side of the image which the measurement was taken for, L = left, R = right & C = centre. Average measurements are calculated for week 1 (Avg1) and week 2 (Avg2) at both near (AvgN) and far (AvgF) distances.

Photographic Session (*Week*) had no significant effect on any of the measurements ($p > .1$ for all), indicating similar viewpoint and expression in both sessions. More importantly for this study, Camera-to-Subject Distance (*Distance*) had a significant effect on some measures but not on others. Specifically, the more peripheral nose-to-mouth measurements were greater for *Far* images than for *Near* images, whereas the more central eye-to-nose measurements were statistically equivalent at the two camera distances we compared. This pattern in the anthropometric data corroborates the flatter appearance of the *Far* images and the more convex appearance of the *Near* images. More generally, it confirms the non-linear effect of camera-to-subject distance on configural information for this image set: some measurements change more than others. I next used a paired matching task to assess the implications of these configural changes for perception of facial identity.

## 4.4 Experiment 8: Face-Matching & Camera-to-Subject Distance

In the GFMT, a standard matching experiment, participants are presented with pairs of face photographs that were taken with different cameras (Burton et al. 2010). For each pair, the participant's task is to decide whether the 2 photos show the same person (*Same* trials; 50% prevalence) or 2 different people (*Different* trials; 50% prevalence). Despite the simplicity of this task, error rates are high when the faces are unfamiliar, as

the viewer has no way to distinguish image changes from identity changes. When the faces are familiar errors are virtually absent, presumably because variation in appearance is better characterised by the viewer (Jenkins & Burton, 2011).

Here I extend the standard paired matching task by adding camera-to-subject distance as an experimental factor. Because face images change with camera-to-subject distance, manipulating this distance allows very specific predictions to be made: for viewers who are *unfamiliar* with the faces concerned, a change in camera-to-subject distance should impair performance on *Same* Identity trials (because it generates dissimilar images) and should *improve* performance on *Different* Identity trials (for the same reason). If identity judgments by familiar viewers rely on facial configurations then the same should apply to their performance. However, given that familiar viewers readily see through changes in viewpoint, lighting, facial expression, and other factors, it is anticipated that familiar viewers might similarly see through changes in camera-to-subject distance, such that their performance would be unaffected by this manipulation

## Method

### Participants

45 psychology undergraduates at the University of York participated in exchange for payment or course credit. 23 of these participants were first-year students who arrived at the University of York after our photographic models had left, and hence had never seen the faces in the stimulus set (verified post-test; see Procedure section below). We refer to these participants as *Unfamiliar* viewers (M = 4, mean age = 18.7). The remaining 22 participants were other students from the same year group as our photographic models, and had spent over two years studying on the same course (M = 3, mean age = 22.14). I refer to these participants as *Familiar* viewers because they had seen the faces in the stimulus set routinely over those two years.

**Stimuli and Design**

The same stimulus images were used as in Experiment 8. Images were cropped around the face to remove extraneous background. The face images measured 700 pixels wide by 900 pixels high.

In order to create the face-matching task, images were paired according to *Same* and *Different Identity* trials. The different identity trials were created by pairing the most similar face images from those available. Image pairs were also constructed according to *Same* and *Different Camera-to-Subject Distance*. Crossing the within subject factors *Identity* and *Distance* resulted in four stimulus conditions: (i) Same Identity, Same Distance; (ii) Same Identity, Different Distance; (iii) Different Identity, Same Distance; and (iv) Different Identity, Different Distance. Figure 4.4 shows example pairs from each condition.

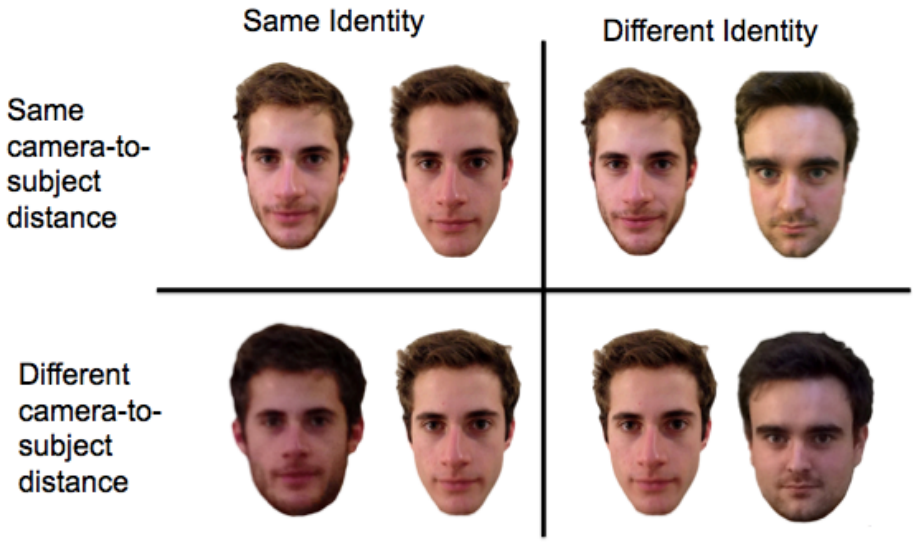


**Figure 4.4** Example of one identity with each of their four image identity pairings shown. The first column shows image pairs of the same identity and the second column shows different identity pairs. The first row shows same camera-to-subject distance pairs and the bottom row shows pairs where the images are of different camera-to-subject distance.

For each pair, the participant's task was to decide whether the two images showed the same person or two different people. This allowed a percentage accuracy score to be calculated for each participant in each condition. Both *Unfamiliar* and *Familiar Viewer* groups carried out this task.

**Procedure**

Participants were tested in groups and worked individually in silence. Participants viewed a total of 144 image pairs as part of a face-matching task. These images were viewed as part of a timed PowerPoint presentation in which each image pair was visible for 5 seconds followed by a 3,2,1 countdown before the next image pair appeared. Within this viewing and countdown time participants were required to decide whether the images in the pair they had just viewed showed the same person's face or if the faces were two different people. Participants recorded their answers by circling 'same' or 'different' on an answer sheet to indicate their identity judgement for each image pair.

After completing the face-matching task, participants completed a familiarity check. For this, participants viewed images of each of the identities that were stimuli in the task and simply ticked a box to indicate whether they were familiar with each face or left the box unmarked if the faces were unknown to them. This allowed me to exclude the data from faces that were unfamiliar to any of the familiar viewer group, or familiar to any of the unfamiliar group.

**Results**

A 2x2x2 mixed ANOVA *between Identity, Viewer Familiarity* and *Distance* was performed and confirmed that overall performance was significantly lower for *Unfamiliar* viewers (M = 85.71, SE = .80, CI = 84.11 – 87.31) than *Familiar* viewers (M = 97.89, SE = .81, CI = 96.25 – 99.53), $F(1,43) = 114.8$, $p<.001$, $\eta_p^2 = .73$, this finding was expected given that familiar viewers have been found to be more robust against superficial image changes than unfamiliar viewers. There was a significant main effect of *Identity*, with performance

being higher for *Different Identity* (M = 93.36, SE = .91, CI = 91.53 – 95.2) than *Same Identity* trials (M= 90.25, SE =.87, CI = 88.49 – 92.00), F(1,43) = 5.17, p = .03, $\eta_p^2$ = .11. However there was no interaction between *Identity* and *Viewer Familiarity* [F(1,43) = 1.53, p =.22, $\eta_p^2$ = .03].

There was also a significant main effect of *Distance*, with participants performing better for *Same Distance* trials (mean = 94.92, SE =.55, CI = 93.82 – 96.03) than *Different Distance* trials (mean = 88.69, SE = .85, CI = 86.97 – 90.38), F(1,43) = 51.90, p<.001, $\eta_p^2$ =.547. This demonstrates that changing camera-to-subject distance lowers face-matching performance accuracy. In addition to this both the interaction between *Distance* and *Viewer Familiarity* [F(1,43) = 21.80, p<.001, $\eta_p^2$ = .34] and the interaction between *Distance* and *Identity* [F(1,43) = 133.64, p<.001, $\eta_p^2$ =.76] were significant. Finally, there was a significant three way interaction between *Identity, Distance* and *Viewer Familiarity* F(1,43) = 117.45, p<.001, $\eta_p^2$ = .73. To break down this 3-way interaction, I next carried out separate 2x2 within-subjects ANOVAs for the *Familiar* and *Unfamiliar* groups.

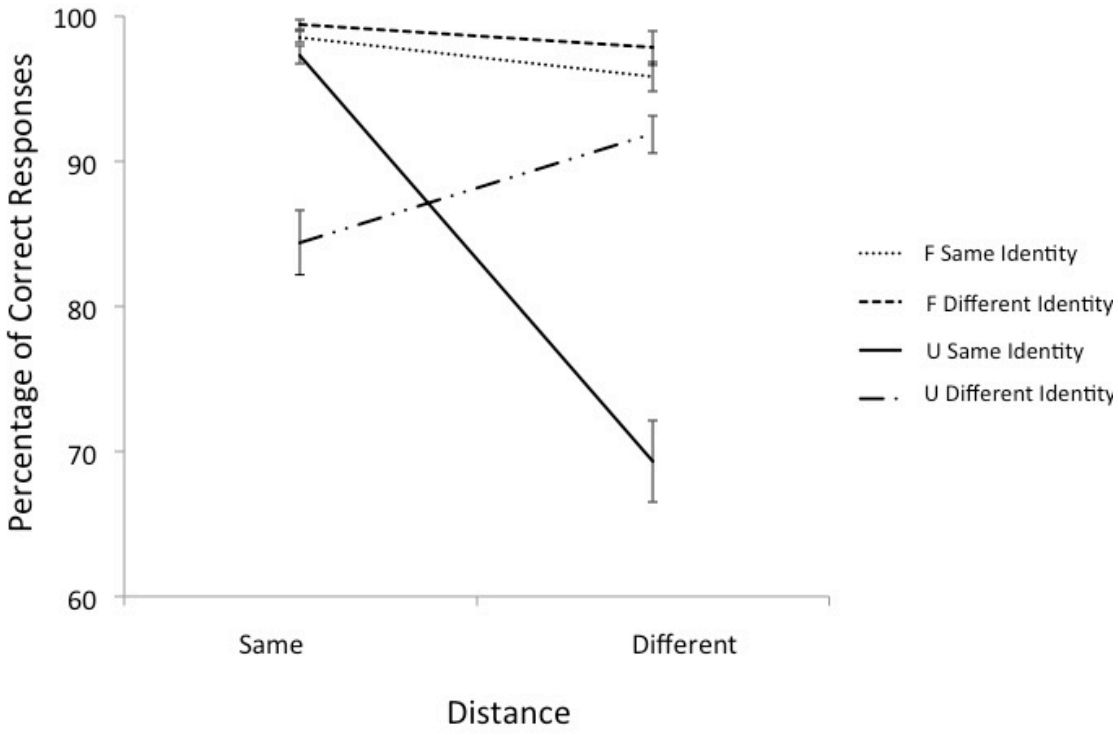


**Figure 4.5** Effect of changing camera-to-subject distance on performance accuracy in the face-matching task for familiar (F) and unfamiliar viewers (U), for same and different identity trials, at both same and different distances. Error bars show the standard error of the mean.

As expected, *Familiar* viewers performed with very high accuracy in both the *Same Distance* and *Different Distance* conditions. Accuracy was significantly higher for *Same Distance* image pairs (M = 98.84 % correct, SD = 1.37) than for *Different Distance* image pairs (M = 96.70 % correct, SD = 4.17), [F(1,21) = 5.49, p<.05, $\eta_p^2$ = .21], even though the magnitude of this effect was small. There was no significant effect of *Identity* [F(1,21) = 3.41, p = .08, $\eta_p^2$ = .14] and no interaction between these two factors p>.05.

More importantly, *Unfamiliar* viewers performed significantly better when the paired images were taken at the *Same Distance* (M = 90.85% correct, SD = 4.97) compared to *Different Distance* (M = 80.54, SD = 6.83) [F(2,22) = 51.20, p<.001, $\eta_p^2$ = .67]. There was no significant effect of *Identity* [F(1,22) = 3.48, p=.08, $\eta_p^2$ = .14], but critically there was a

significant *Identity Distance* interaction [F(1,22) = 155.13, p<.001, $\eta_p^2$ = .88]. For *Same Identity* pairs, accuracy was higher for *Same Distance* (M = 97.33, SD = 3.09) images than *Different Distance* (M = 69.29, SD = 13.41) images. For *Different Identity* pairs the opposite was true: accuracy was higher for *Different Distance* (M = 91.86, SD = 6.29) images than for *Same Distance* images (M = 84.37, SD = 10.64). Simple main effects revealed that there was a significant effect of *Identity* for both *Same* [F(1,44) = 19.34, p<.001, $\eta_p^2$ =. 31] and *Different Distance* images [F(1,44) = 58.73, p<.001, $\eta_p^2$ =. 57]. And there was a significant effect of distance for *Same Identity* pairs [F(1,44) = 191.92, p<.001, $\eta_p^2$ =. 81], and also for *Different Identity* pairs [F(1,44) = 13.69, p<.001, $\eta_p^2$ =. 31].

**Discussion**

Changing camera-to-subject distance impaired viewers' face-matching ability. Accurate face recognition remained easy for viewers who were familiar with the faces concerned regardless of camera-to-subject distance (mean performance across same distance and different distance conditions = 99% accuracy). Unfamiliar viewers performed much more poorly in the different distance condition (M = 81%) than the same distance condition (M = 91%). There was a crucial cross over interaction in the unfamiliar group – with performance decreasing as a result of changes to camera to subject distance only for same identity trials.

This pattern of results for unfamiliar viewers suggests that the kinds of configural change evident in the measurement study join the long list of image changes that can thwart identification. This finding shows that the effects of changing camera-to-subject distance identified in the past (Latto & Harper, 2007) do indeed also impair face recognition. Distance related image changes had been found to impair face memory for digitally edited compared with identical images (Liu, 2003). My study demonstrates that distance manipulations also impair face-matching.

The finding that familiar viewers were impervious to these non-linear changes in facial configuration suggests that familiar face recognition is not strongly dependent on distances between features in face recognition. It seems that when learning a face, people are not learning any specific configurations of a face, as familiar viewers are able to see through changes in camera-to-subject distance, where these configurations differ across the images.

The findings of these studies raise an interesting question: if faces look more flat or convex at different viewing distances, why do we not notice these shape changes in daily life? The next experiment addresses this question.

## 4.5 Experiment 9 - Perceptual Constancy for Face Shape

In the measurement study I demonstrated that images undergo non-linear changes in configuration as a result of changed distance. Experiment 8 showed that these image changes can easily disrupt unfamiliar face matching: unfamiliar viewers were poorer at matching pairs of faces when the two images were taken from different camera-to-subject distances, compared with when the images were taken from the same distance. In real life however, we do not tend to notice changes in face shape. For example, faces of people walking towards us do not appear more convex as they approach. There are in fact lots of examples of not noticing image change for images other than faces. These include changes in colour and brightness, for example clothes are perceived as the same colour even under different types of lighting; the colour is perceived as constant through calibration of white (Webster & Mollen, 1995). Shapes are also perceived constant through the use of depth cues (Pizlo, 1994). Each of these scenarios is an example of perceptual constancy – where the visual system uses information from the environment to overcome image changes. Perceptual constancy has been studied intensively but normally in the context of low-level visual features such as colour and shape.

It is evident that perceptual constancy can help viewers make sense of the colours or shapes that they see. Following this, one possibility is that perceptual constancy mechanisms compensate for changes in face shape caused by viewing distance. In the same way that viewers perceive the shape of an open door as rectangular, even though the retinal projection is trapezoidal, they may view face shapes as a constant shape despite distance-related distortions by taking account of the viewing distance for the face image. It may be that when looking at a photograph on a screen no information of viewing distance is available, removing the ability to compensate for distance and as a result the face images look different.

In order to investigate if perceptual constancy applies to face shape under these conditions I will investigate whether a distance cue, indicating that one image was taken from further away and one close up, can overcome the effect that a change of camera-to-subject had on recognition accuracy as found in Experiment 8. My approach to this will be to manipulate congruency between the actual camera-to-subject distance and the distance implied by cues in the display. Specifically, I will present the two face images in each pair at two different sizes – a small image implying a long viewing distance and a large image implying a short viewing distance. In congruent trials, the size cues will be in sympathy with the images, so that the near image is larger and the far image is small. In incongruent trials, the size cues will conflict with the images, so that the near image is small and the far image is large. If a perceptual constancy mechanism compensates for distance-related distortions, then participants should perform more accurately on congruent trials than on incongruent trials, when the two images show the same person. This is because the valid distance cues will allow constancy to correctly 'undo' the optical distortion, making the images look more similar. At the same time, accuracy should be lower for congruent trials than for incongruent trials, when the two images show different people. This is because with invalid distance cues, the constancy mechanism will compensate in the wrong direction, exaggerating differences between the images, making them look less similar.

## Method

### Participants

30 unfamiliar participants (male=6, age = 19) from the University of York took part as participants in this study.

### Design

Experiment 9 was largely based on the design of Experiment 8. However, in this experiment the image pairs featured a distance cue. In half of the trials this distance cue was *Congruent* with the true distances for which the photos were taken e.g. the near photo was shown as a bigger image on screen and the far image as smaller. Other times I swapped the image sizes, to create an *Incongruent* image pairs e.g. the near image was made smaller to appear far away, and the far image enlarged to appear near.

*Congruent* image pairs were created by keeping the near photograph its natural size (the original size of the saved photograph, not resized in any way), and the far image its natural size. This resulted in a larger 'near' photograph, and a smaller 'far' photograph. *Incongruent* images were created by resizing the far image to be the natural size that the near image was, and resizing the near image to take the size that the original far image was. This created a small 'near' image and a large 'far' image, hence incongruent to the natural display format these images would take. In addition to this, perspective lines were added to the images to provide an additional depth cue. The perspective lines supported the interpretation of distance in the displays. See Figure 4.6 for an example of congruent and incongruent same and different identity pairs.

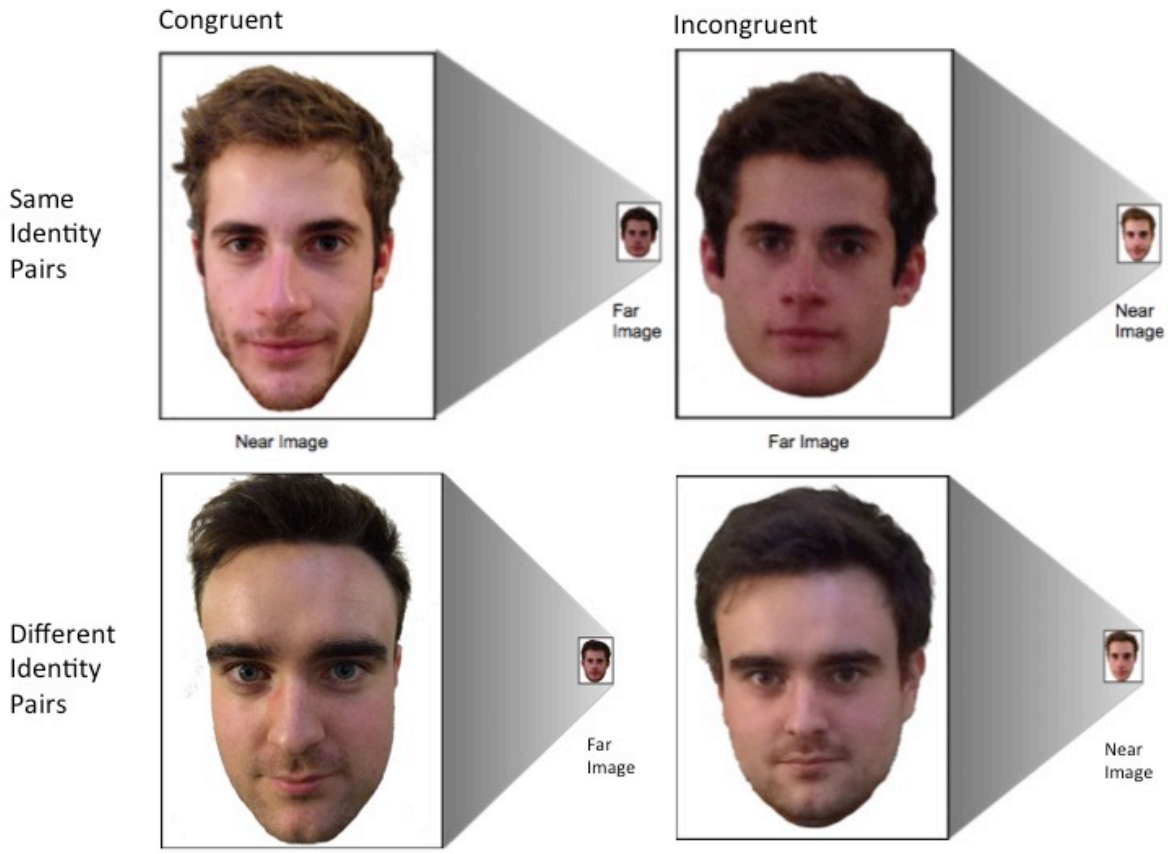


**Figure 4.6** Example of congruent and incongruent face image pairs (with distance cues) for same and different identities.

### Procedure

The task was a face-matching task as in Experiment 8, however the stimuli now featured distance cues. Participants were tasked with making same or different identity judgments for each face pair that they saw. 144 pairs were viewed in total.

**Results & Discussion**

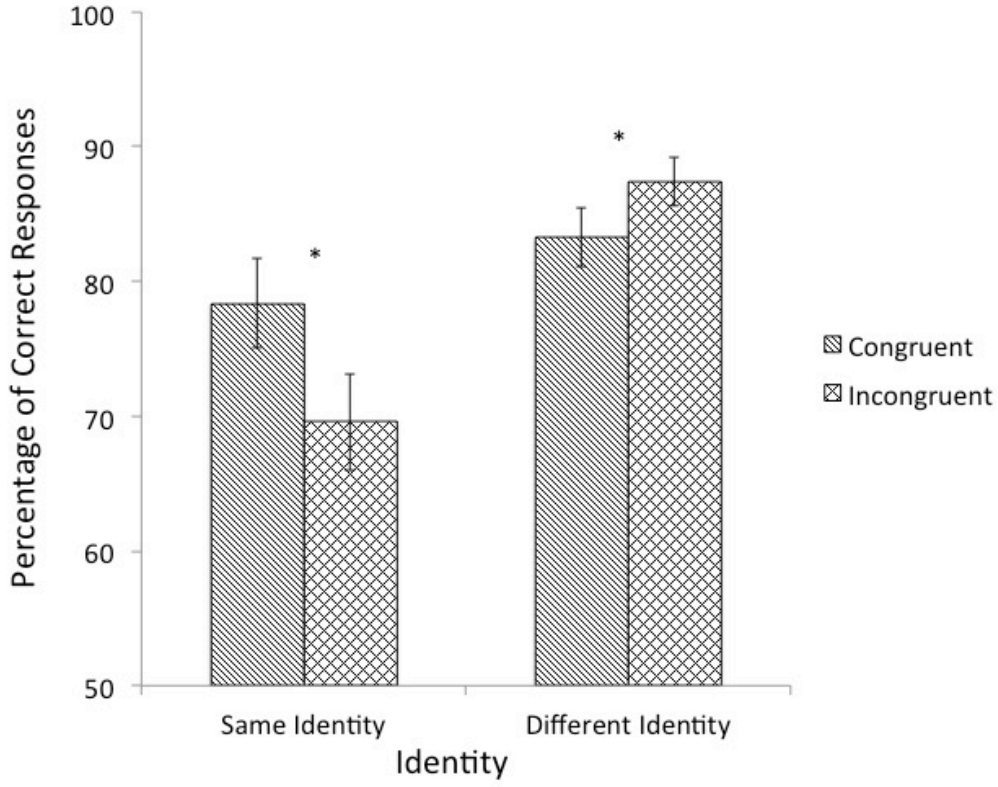


**Figure 4.7** Graph showing the percentage of correct responses for both congruent and incongruent image pairs broken down by same and different identity trials. Error bars show the standard error of the mean.

A 2x2 ANOVA was conducted to compare the effect of *Distance Cues* (*Congruent* & *Incongruent)* and *Identity* (*Same* & *Different*), on face matching accuracy (see Figure 4.7). The analysis revealed no significant main effect of *Distance Cue*, comparing accuracy on *Congruent* (M = 80.82) with *Incongruent* (M = 78.49) trials F(1,27) = 3.29, p = .08, $\eta_p^2$ = .11. However, there was a significant main effect of *Identity* F(1,27) = 8.13, p = .008, $\eta_p^2$ = .23, with higher recognition accuracy for Different (M = 85.35) rather than Same (M = 73.97) identity trials. More importantly, there was also a significant interaction between *Congruence* and *Distance*, F(1,27) = 23.22, p<.001, $\eta_p^2$ =.46, confirming that the congruency manipulation had opposite effects for *Same Identity* and *Different Identity* trials.

Simple main effects revealed that as anticipated, participants were significantly more accurate at *S*ame identity trials for *Congruent* image pairs (M = 78.38, SE = 3.3) than *Incongruent* image pairs (M = 69.57, SE = 3.6), F(1,54)= 22.44, p<.001, $\eta_p^2$ = .29.

My second prediction was also met with results showing that for different identity trials, viewers were more accurate for incongruent image pairs (M = 87.42, SE = 1.8), than for congruent image pairs (M = 83.27, SE = 2.18), F(1,54) =4.97, p<.01, $\eta_p^2$ = .08.

The results from Experiment 9 suggest the operation of a perceptual constancy mechanism at the level of face shape. For unfamiliar faces, participants were better for same identity trials when provided with a congruent distance cue. In this situation the constancy mechanism would be giving the information necessary to make the images look more similar e.g. compensating in the correct direction for the differences in images as a result of distance.

Further evidence for the constancy mechanism comes from the result found for different identity trials. In the case of incongruent different identity face trials if the logic applied above was followed, the incongruent distance cues would lead to being compensated in the 'wrong' direction, making the images look even more different than they would otherwise. Hence correct different judgments were greater when the differences were amplified due to incongruence than when they were viewed in the congruent condition. If no perceptual constancy mechanism were being applied to face perception no differences in performance would have been observed across conditions.

## 4.6 General Discussion

In this chapter I have shown that changes in camera-to-subject distance lead to non-linear changes in face measurements across images. These changes had a detrimental effect on face identification efforts of unfamiliar viewers – identity matching performance was much poorer when unfamiliar viewers compared photographs taken from differing distances, than when the comparison images were taken at the same distance. This effect was driven by difficulty for matching same identity trials. Familiar viewers were far less affected by this distance change and performed at very high accuracy levels in both conditions. Furthermore I provide evidence of perceptual constancy effects at the level of face shape. Congruent distance cues aided recognition of same identity faces, with distance cues compensating for the apparent differences in faces due to camera-to-subject distance. In line with this, an incongruent image cue made the images look even more different to each other, and hence increased accuracy on correct different person judgements. These findings suggest that perceptual constancy can account for the apparent continuity of facial appearance at different viewing distances, and that valid distance cues are required for the smooth operation of this mechanism.

Previous research had shown that perceptions gathered from a face image, including weight estimates and social inferences, differed as a result of changed camera-to-subject distance (Harper & Latto, 2001; Bryan et al. 2012). My analysis showed that non-linear changes in metric distances between facial features can account for such perceptions. Most importantly, in addition to the social inference effects reported previously, my study demonstrated that these configural changes affect accuracy in *identity* judgment. The camera-to-subject manipulation greatly impaired unfamiliar viewers, whereas familiar viewers were barely familiar at all. The level of familiarity with each face was not recorded – participants rated faces as either familiar or unfamiliar. It could be the case that some of the familiar viewers were not all that familiar with the faces involved, which could explain the small difference found for familiar viewers over distance change. My findings are very much in line with past research demonstrating superior performance of familiar viewers in face recognition task over unfamiliar viewers (e.g. Burton et al. 1999,

Jenkins et al. 2011). In addition this research is the first to show that face matching performance, in addition to face memory (Liu, 2003) is impaired by camera-to-subject distance change across images.

Moreover, the findings of this chapter suggest that learning configural information – at least in the conventional sense of distances between features – is not the key to learning a face. Experiment 7 highlighted that the configuration of a face does not remain constant across multiple images if these images have been taken from different camera-to-subject distances. Yet, it is known that people can accurately identify celebrities who have become familiar faces through our exposure to images and video footage, which capture the celebrity from many different distances. It seems more likely that when people are become familiar with a face they are gaining experience under a variety of conditions, this would include learning how the face looks from a range of different distances (e.g. Jenkins, 2011). This fits with the conception of face space advanced throughout this thesis. On this account, familiar viewers have a range of experience with a face, and hence experience of seeing the face over several different distance configurations. Unfamiliar viewers do not have this refined face space and must instead make identity judgments using only the information that is in the images in front of them. This lack of perceptual constancy interacts with configural change in the following ways. First it makes it harder to 'tell together' same identity images with different configurations, as the viewer does not have experience of the variability in appearance for that identity. Second, it supports viewers' performance on different identity trials, as the different camera-to-subject distances tend to exaggerate natural differences between the faces such that the identities would be more likely to be categorised as belonging to different identities in face space. Notably, the distance change only makes different people look more different because they look similar to begin with - it would be possible for different people to look less different as a result of distance change if for example one was fat and one was thin.

These experiments also advance the theoretical understanding of unfamiliar face recognition; in particular I have proposed that a perceptual constancy mechanism exists whereby distance cues help an individual make sense of a face image viewed. Whereas past studies have shown perceptual constancy for colour and basic geometric shapes, and even for the size and shape of humans, to my knowledge, this is the first study to propose a constancy mechanism for faces.

Several practical implications follow from these findings. Firstly, according to the results of Experiment 7, anthropometry (in which the images used in investigation are likely taken from unknown different distances) is not a reliable method of facial identification. This is because images of faces do not hold constant configurations - measurements between features of a face change across images when these images have been taken from different distances (Kleinberg, Vanezis & Burton, 2007). Additionally, as the results of Experiment 8 show that changing the distance from which a photograph of a subject was taken between images reduces identification accuracy when making identity judgments based on face-matching, it would be advised that wherever possible, there should be consistency in distance from which photographs are taken from in forensic or security situations. For example better identification rates may be met if security officials took photographs of a suspect from a standard distance, which could be compared with photographs taken of the same suspect, from this same standard distance, following a separate incident, no matter where the incident occurred. If this is not possible, caution is urged when comparing images that were taken at different or unknown distances, because there are systematic differences in how they will appear. Familiar viewers are largely exempt from this caution because they can see through configural changes easily. Where this is not possible, providing accurate distance cues could improve performance for unfamiliar viewers.

In conclusion, face configuration in an image changes as a result of changes in camera-to-subject distance. These changes affect performance in face-matching tasks, with unfamiliar viewers being very strongly affected. The finding that accurate distance cues

help viewers to see through these changes in configuration suggests the operation of a perceptual constancy mechanism.

Up until now this thesis has addressed incidental creations of, or naturally occurring, challenging stimuli. The next chapters of this thesis will move on to explore a second type of challenging stimuli, where deliberate attempts are made to change appearance through disguise.

# Chapter 5 – Matching Disguised Faces

## 5.1 Chapter Summary

I created a sophisticated disguise face database which is the first to include both evasion and impersonation disguise items. Models were unguided in their disguise efforts. Any props that they requested to aid their disguises were provided and all models were informed of the effect of changing face shape as a result of camera to subject distance (Chapter 4).

Disguise impaired face-matching performance. Performance for unfamiliar viewers was poor, and knowing that disguises may be present did not improve this. Evasion disguise caused most difficulties for face-matching, followed by 'impersonation similar' then 'impersonation random' disguises. Familiar viewers were much better at seeing through disguise than unfamiliar viewers but even familiarity did not help viewers completely see through evasion disguise. Links to theories of face space and familiarity are discussed.

## 5.2 Introduction

This chapter will explore face-matching performance for disguised faces. Previous research has shown that unfamiliar face-matching is poor, with people making between 10-20% errors on the GFMT (long & short version respectively) which is a standardised test of face image matching, and also in matching tasks involving an ID card image and a physically present person (Burton et al. 2010; Kemp et al., 1997; White et al., 2014). However these results come from tasks that have used cooperative stimuli (when same identity trials were created there was no deliberate attempt to make an individual look different across multiple images) and mismatch trials were created from the most similar match taken from a limited sample of face images.

I found that performance was even worse when using more difficult stimuli images involving incidental disguise (Chapters 2 & 4) than had been found previously for cooperative stimuli. These include the case of ambient same person images and extremely similar imposter face images (Chapter 2) and images where face shape is changed as a result of camera-to-subject distance (Chapter 4). These studies have helped in gaining a better understanding of face recognition in challenging circumstances and my findings have also demonstrated the power of familiarity – so far familiarity has led to large increases in performance on the difficult tasks that I have developed.

With such promising results, I now invest in creating a disguise face database to allow face-matching performance accuracy to be tested in what may be the most challenging case yet – when a person is *deliberately* trying to evade their own identity or impersonate someone else. As I am creating the database from scratch it will include all aspects of disguise that I am interested in. Past studies have relied on incidentally occurring similarities between faces to provide different person trials in face-matching tasks and used two images of the same person that may differ in naturalistic incidental ways. Now, in addition to creating no disguise image pairs similar to those used in past matching tasks, I attempt to create disguise image pairs. These will include creating deliberate similarities between different identities (impersonation disguise) as well as creating same person trials where there is a deliberate attempt to alter appearance across images (evasion disguise). These images will allow face-matching performance for disguised faces to be directly compared to performance for undisguised faces for the first time.

**Disguise in Face Recognition Memory**

Only a few past studies have attempted to approach questions related to face recognition performance and disguise. Patterson and Baddeley (1977) were the first to do so, publishing a paper that consisted of two disguise face memory experiments. Their first experiment tested face recognition memory performance for identical face images compared with face images changed in pose or expression, and also faces in disguise. The

disguise stimuli were images of actors, with the same actors photographed in two different character roles. This allowed one of these images to be presented during the learning phase and the other at test. Participants were divided into two groups for the learning phase in the experiment. Both groups were required to view face images with the intent of memorising them. However, one group was instructed to focus on making personality judgments for the faces they saw and the other instructed to focus on the features of the faces (discussed further in Chapter 5). Participants were aware that the appearance of the individuals might change when they saw the face again at test. At test participants again viewed face images, and were this time tasked with deciding whether or not the identities presented at test had featured in the earlier memory task. Recognition performance was poorer for images presented in disguise at test than images that were presented unchanged [identical image presented again at test] or changed only in pose or expression. Recognition performance was near chance for faces presented in disguise at test.

The disguise stimuli used in Patterson & Baddeley's (1977) first experiment in their paper discussed above, were multiple images of the same actors, whose appearances differed across images as a result of matching different character roles for different acting jobs. This stimuli acquisition method meant that the exact changes made to appearance by the actors across images were unknown to the experimenters, and hence the effect of each change could not be easily investigated. A more controlled manipulation of disguise was applied in Experiment 2. Here Patterson & Baddeley created the disguise stimuli themselves by asking volunteer male models to create disguises using props. Disguise through purely the addition of props is a common approach to disguise undertaken by past work. I will here on refer to disguise by props alone as *simple disguise*. Models in Patterson & Baddeley's study were instructed to add wigs, add or remove beards, and add or remove glasses across a series of photographs. I will refer to disguise based on the experimenter's instruction as *prescribed disguise* and situations where models disguise themselves as they wish as *free disguise*. Patterson & Baddeley's experiment thus fits the category of prescribed simple disguise. Participants were tested on their recognition performance for disguise compared to no disguise images in a similar manner to in

Patterson & Baddeley's Experiment 1. As in the case of the recognising disguise actor images, participants in Patterson & Baddeley's (1977) second experiment were worse at recognising a face identity when it was presented in a different disguise at test to learning, than recognising an exact image that they had viewed before. Specific effects of each disguise manipulation investigated in this experiment are discussed in Chapter 5.

Terry (1994) also explored the effect of disguising individuals on memory for computer-generated faces. Experiments included the addition or removal of glasses, or the addition or removal of a beard on computer generated face images. Removing glasses reduced recognition accuracy, as did the addition of a beard, and also the removal of a beard, but to a lesser extent. However, adding glasses did not affect recognition accuracy. Righi, Peissig & Tarr (2012) replicated this result with real face images from the TarrLab database, showing that the addition of glasses did not affect face recognition memory performance as much as the addition of a wig or the removal of glasses. These findings show that a change in appearance in form of a disguise is not always clear cut in terms of the effect it will have on face recognition – some manipulations impair performance whereas other do not.

Both of these studies (Patterson & Baddeley 1977 & Terry, 1994) report overall impairments to face recognition memory when an image is presented in a changed form (in a disguise) at test compared to at presentation despite significant difference in the disguise stimuli used in each experiment. Patterson & Baddeley's (1977) Experiment 2 and Righi et al. (2012) consisted of prescribed disguise stimuli whereas Terry (1994) used computer-generated stimuli images. This prominent effect of disguise on face memory makes it seem likely that face-matching tasks will also be affected by disguise manipulations as factors such as changes in pose and expression, which have influenced face recognition memory, have also reduced performance for face-matching in past studies (Bruce, 1982; Hancock et al. 2000).

**Disguise in Face-matching**

Dhamecha, Singh, Vatsa & Kumar (2014) are the only researchers to have tested human face-matching accuracy for disguised faces. Their study compares human face recognition performance to machine algorithm performance. The main reason for human performance being investigated was to learn from the strategies used by humans in order to enhance the algorithms.

**Figure 5.1** Images from the IDV1 database. Props include glasses, fake beards and moustaches, medical maskes and hats turbans.

Dhamecha et al. (2014) created the IIII-Delhi Disguise Version 1 face database (IDV1) for their study (see Figure 5.1). This consisted of images of 75 models each of whom had been given access to accessories that they could use to disguise themselves. Models were photographed in 5 disguises, and in 1 no disguise (neutral) condition. All photographs were taken under nearly identical lighting conditions, with a frontal pose and neutral expression, thereby limiting options for creating disguise. Accessories for disguise were wigs, fake beards and moustaches, sunglasses and glasses, hats, turbans, veils, and medical masks. Models could wear just one, or multiple accessories for their disguises but were simply told 'to use the accessories at their will in order to get disguised', therefore the models conducted simple free disguise. The props and the study itself focus towards occluding specific areas of the face (e.g. eyes hidden by glasses, medical masks hiding mouth). Occlusion of features is a typical result of many simple disguise manipulations. The authors point out that the photographs used as stimuli included all parts of the face being hidden at least once. Although occlusion of features may be effective in hiding identity, it is important to remember that concealing parts of a face would often not be an effective disguise in the context of identity fraud as these disguise props would have to be removed for facial comparison inspections. Concealment of features through use of

props is a different type of disguise to manipulating the way features themselves look and may lead to very different results.

The study reported that human performance accuracy for matching the face pairs differed as a result of ethnicity of the faces viewed and also familiarity with the faces. Performance accuracy was lowest for matching unfamiliar faces of different ethnicity (66% accuracy), followed by unfamiliar faces of the same ethnicity (69.5%) and highest for the familiar faces, all of which were the same ethnicity (75% accuracy). Performance of the algorithm proposed in the paper was comparable with the unfamiliar, different ethnicity human viewers (Dhamecha et al., 2014).

Dhamecha et al. (2014) provide the first assessment of disguise face-matching accuracy by human viewers. In particular, the free disguise aspect used by Dhamecha and colleagues was interesting in that it produced a stimuli set of faces that differed from each other in their disguises, with multiple disguises for each face. However, this design also had limitations. One specific area of concern is that there were inconsistencies in the matching task. Participants in the study made same or different identity judgments to pairs of face images. Sometimes this involved matching a neutral face to a disguise face, but more often participants were matching pairs of disguised images. Nearly all the faces had some type of disguise present, therefore the study compared performance between familiarity and ethnic groups on a task of matching disguised faces rather than investigating whether disguise impaired performance more than the matching of undisguised face pairs. This design meant that a direct comparison between performance accuracy for disguised and undisguised faces could not be made. An additional limitation of the design is that images were cropped so that only internal cues could be used. This disregards any influence of external cues which may actually have been able to help contribute, or alternatively could have worsened, human participants' face-matching performance. This would have been an interesting comparison to have, particularly as previous work suggests that unfamiliar viewers rely heavily on external features when making facial identity judgments (see Chapter 6 for further discussion of features and

identity judgments) (Bruce et al, 1999, 2001; Bonner, Burton & Bruce, 2003; Megreya & Bindemann, 2009).

**Limitations of Previous Disguise Research**

There are four main limitations of the previous research conducted on facial identity performance for disguised faces. These are i) that past tasks compare recognition of disguised faces with identical image comparisons, ii) a focus on memory tasks, iii) investigation of only evasion disguise, ignoring impersonation, and iv) the use of simple, often prescribed, disguises as the stimuli. The reasons why each of these points limit disguise research is discussed below.

Firstly, performance for disguised face identity judgments have most often been compared with performance for remembering exact images. Patterson & Baddeley, (1977) forced comparisons of image recognition (remembering the exact image) with a difficult case of face recognition (different images of the same face), rather than testing whether disguised face recognition is more difficult than undisguised face recognition (across different undisguised images of the same identity), which would be a more meaningful comparison. Some of the studies have actually made alterations to the learning image itself to create a disguise condition, which recreated a scenario more similar to image matching than purely investigating disguise (e.g. Terry 1994).

Recognition of identical images is an easy task for humans (Bruce, 1982). Facial appearance changes incidentally across multiple different images. In real life face-matching scenarios, it is performance across these types of changes that are of interest, as it not possible to capture the exact same image of a person across time. This has now been recognised in the study of face recognition. Ambient face images, which include naturalistic and incidental changes, such as differences in expression, are now typically the image type used in tasks that try to capture facial identity performance accuracy.

Comparisons with image matching performance do not match on to the real world problem where naturalistic differences exist between multiple images of a face. In the investigation of disguise, it is thus important that performance for disguised faces is compared against performance for faces that include this incidental change in appearance

Further to the focus on image matching, past research is mostly confined to the study of face recognition *memory* for disguised faces (Patterson & Baddeley 1977; Terry, 1994; Righi et al. 2012). This thesis focuses on face-matching accuracy, as this is a frequently used identification check that does not rely on any memory component. The only task to date which has looked at face-matching accuracy for disguise faces (Dhamecha et al. 2014) investigated the performance of computer algorithms on the task compared to human performance, but did not provide a control of performance for faces in no disguise. Previous research has found that unfamiliar face-matching is poor in cooperative facial image comparison task, but I believe it to be interesting to know whether disguise impairs face recognition performance further. In criminal or undercover police situations there can be very strong incentives to carry out a realistic and successful disguise, but it is not yet confirmed that disguises impair face-matching performance.

Previous research on disguise has focused exclusively on evasion – changes to appearance that make the model's own identity difficult to determine. Disguise could also involve impersonation, changing appearance to look like a specific other person, but this has not been addressed by disguise research to date. Most stimuli databases include exclusively male faces and some studies have relied on computer generated stimuli rather than real human face images to explore evasion disguise. A database of male and female human faces in both evasion and impersonation disguise is necessary to better understand disguise.

A final limitation of the previous research on disguise is that the stimuli images used have been unsophisticated. Simple disguises have dominated which involve the addition or removal of props to a face from a limited supply of props provided by the experimenters. This limits the ways in which individuals can disguise themselves and may not reflect the props that the person would choose to disguise themselves naturally. Simple disguise has often led to occlusion of features. This disguise technique would create ineffective disguises in many contexts of identity fraud. Most disguises have been prescribed by the experiments, meaning that no information has been gathered on the changes that people naturally make to their appearance to create disguises. Patterson & Baddeley (Experiment 1) were the only researchers to test more naturalistic disguises in terms of the actor photographs, but the intent at time of photograph for these appearance alterations was not specifically disguise. Examples of existing disguise face databases are shown below - figures make limitations of the stimuli involved evident.

## 5.3 Existing Disguise Face Databases

**The AR & IDV1 Databases**

The AR database (Martinez & Benavente, 1998) and the IDV1 database (Dhamecha et al., 2014) are the only existing disguise face databases to contain images of real people under a variety of disguises. In the case of the AR database, disguises include changes in lighting and deliberate changes in expression, which led to incidental changes in the appearance of the identity, and also more deliberate changes to appearance through the addition of props – sunglasses and a scarf to hide features of the face (Figure 5.2). This database consists of only changes that constitute evasion disguise and incidental appearance change. The IDV1 database (image shown previously in Figure 5.1) kept expression and lighting constant; therefore disguise was created purely through the addition of props. Both the AR database and the IDV1 database show rather unconvincing disguises, in that it is obvious that people are trying to hide their appearance through props.

**Figure 5.2** Images from the AR database. Disguise manipulations are limited to a change of expression or the addition of sunglasses or a scarf.

### The National Geographic Database

The National Geographic Database (Ramathan, Chowdhury & Chepella (2004) also includes real faces image with a disguise component, but consists of 46 images of just one identity. It is questionable whether this database should really be considered a disguise database at all, as many of the photographs (e.g. the top row in Figure 5.3) fit the category of incidental change as they show natural variation between multiple images of a face rather than a deliberate attempt to change appearance. The bottom row (Figure 5.3) does consitute disguise images, but these disguises are very obviously fake. It appears that the disguises have been applied to the images themselves (e.g. addition of beard or moustache) rather than to the model prior to the photograph being taken.

**Figure 5.3** A sample of images from the National Geographic Database (Ramathan, et al., 2004).

### TarrLab Face Database

A final example of human disguise face images used in the previous research is the TarrLab face database. This includes images of faces photographed naturally and also the same identities photographed wearing a wig and glasses (Figure 5.4). Sometimes these glasses are reading glasses and at other times they are sunglasses, which occlude the model's eyes. An advantage of the TarrLab database is that images are of both males and female faces, whereas all other databases discussed consist of male faces only. Additionally, the fact that models are photographed both with and without disguise allows direct comparisons of disguise to be made. The database is however limited to prescribed simple disguise, and many of the disguises include occlusion of features. As discussed previously, items that occlude features may have to be removed in security screening scenarios.

**Figure 5.4** Example images taken from the TarrLab face database.

### Synthetic Face Database

Disguise face databases made up of images of human faces have been rather limited to date. Singh, Vatsa & Noore (2009) acknowledge and address this by creating a new disguise face database which encompass a greater range of disguises, but are not images of real human faces (Figure 5.5). Singh et al. (2009) used Faces Software, which is a programme used to make face images based on facial descriptions in police investigations, to build synthetic faces in different disguise conditions. Computer algorithm performance was then tested on this database, with the best performing algorithm performing with 71% accuracy when dealing with images that had multiple components of disguise. The study looked exclusively at algorithm performance on the disguise face database, with the aim of identifying which algorithm performed with highest accuracy on this image set. There was no control condition to compare performance for undisguised face images. This study is therefore limited to evasion disguise performance for computer algorithms on a set of computer generated disguised face images. Moreover the disguises are not realistic, for example, a pirate style eye patch is unlikely to be worn by a disguise perpetrator.

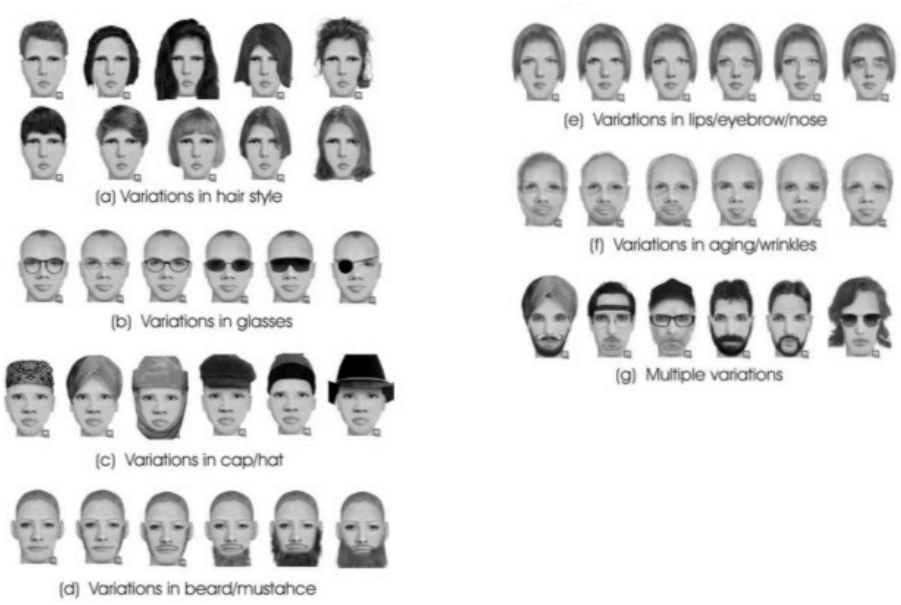


**Figure 5.5** Examples from the synthetic face disguise database (Singh et al. 2009).

## 5.4 FAÇADE Database

### Shortcomings of Existing Databases

As seen from previous research, existing databases are limited - they are unrealistic or involve occlusion of features and include only evasion disguise images. It is also extremely important to investigate impersonation disguise, as in some instances people may have very strong incentives to pass themselves off as a specific other person. Impersonation could include the case of illegally accessing a country on a stolen passport.

Furthermore, past databases are rather unsophisticated in the disguise manipulations applied. All of the disguise databases rely on simple disguise. Many are prescribed disguises (whereby the experimenters have told the model's exactly how to apply the props) but in cases where models have been free to disguise themselves as they wished disguises were severely restricted to the addition of a limited selection of props and combinations of these. As models were never given the opportunity to disguise

themselves with free reign over changes to their appearance, very little is known about the disguise process itself, let alone, what works.

With these shortcomings in mind, I created my own database of disguised and undisguised faces, named the FAÇADE image database, which I will use to investigate deliberate disguise.

**Goals for the FAÇADE Database**

I created the FAÇADE image database to allow face recognition to be explored more fully. Creating a disguise database from scratch allowed the effect of disguise manipulations and the process in creating disguises to be investigated in more detail than ever before. A major goal of this database was to allow research to be conducted to directly compare face-matching performance for disguised against undisguised faces. In addition to this I wanted to break down disguise further – to explore evasion disguise and compare and contrast this to impersonation disguise for the first time.

Unlike past disguise databases, which were made up of images of models who had applied prescribed simple disguises, I wanted to be able to explore how people disguised themselves when they were given the opportunity to create their own disguises. Use of free disguise will allow insight into what measures people would most naturally take to disguise themselves and also allow me to explore what disguises in the database worked best based on results from the matching task. This exploratory analysis will be conducted in Chapter 6.

My initial goal and focus, following the theme of face-matching performance in challenging situations investigated throughout the rest of this thesis, was to assess whether face-matching performance is influenced by the presence of disguise.

**Approach to the Image Acquisition**

As mentioned above, I distinguish between two different types of disguise within my disguise stimuli—Evasion (trying not to look like oneself) and Impersonation (trying to look like a specific target person). To capture this distinction, I gave 26 volunteer models (i) a photo of themselves, and asked them to make themselves look *unlike* that reference photo for a subsequent shoot (Evasion condition), and (ii) a photo of someone else, and asked them to make themselves look *like* that person (Impersonation condition). It is possible that similarity of the impersonation face may influence disguise effectiveness. To investigate this, two impersonation conditions were present for each identity – impersonating a similar face (rated to be the most similar out of 33 match faces by a group of 3 viewers) and impersonating a face that was selected at random. A no disguise photograph was also taken of each model.

Models were instructed that the task was to look as much like or unlike the identity of the target profile photograph, rather than the image itself i.e. I was interested in an identity match task rather than image matching, which itself has already been addressed as a shortcoming of previous research.

One of my goals for the FAÇADE database was to create a database that would not only consist of realistic disguise stimuli for both undisguised, evasion disguise and impersonation disguise faces, but in the acquisition of which, could further the understanding of the things which people do to create each of these disguises. Models were thereby unguided in their disguise effort, as it was of interest to us to find out what people did to disguise themselves. To allow this exploration of changes made, models wrote down all changes they made to their appearance, and the intent of each change, as they carried out their disguise transformations.

I provided models with any props that they wanted to aid their disguises. Props were ordered on request to match the models wishes, but including clothing, wigs, hair accessories, makeup, glasses and jewellery (see Figure 5.6 for selection of props used). Hats, sunglasses and any other occluding items were not allowed, as these props have to be removed if someone was physically present in these items at passport security. Models were instructed that their resulting appearance must not be considered out of place as being a real person's work identity badge image, rather than fancy dress.

**Figure 5.6** Sample of props used to create the disguise face database.

Photographs were taken at a time that suited the model, with photographs (evasion, impersonation similar, impersonation random) not necessarily captured on the same day. This was left to personal choice of the model and gave our models time to make changes to their appearance between photography sessions such as waiting for hair and beard growth.

**Motivation**

It was important that our models were highly motivated and dedicated to create convincing disguises. This was a very time consuming task for our models and required

much thought and some made large changes to their appearance including having their hair cut or died, or beard shaven/grown. It was up to the models which changes they made, so they could decide how far they wanted to go with the disguises. Models were motivated by cash reward, with the best disguise (based on performance in Experiment 1) in each condition (evasion, impersonation similar, impersonation random) receiving a cash reward of £50. The models were extremely dedicated to the task and became competitive with one another to create the best disguise.

**Selecting the best disguise images**

Several images (9 images) of the models were taken in each disguise condition, with models varying aspects such as facial expression, pose and lighting to try and create the best image for each disguise condition (see Figure 5.7). A group of four unfamiliar viewers (the stimuli selection group) worked together to decide upon the most convincing match or mismatch images for each model in each condition. The stimuli selection group viewed the *Impersonation* (similar and random) images and the *Evasion* images for each of the 26 models alongside the corresponding reference image of the target face (the model's own face in evasion condition, and the faces of the people the models were trying to impersonate in the impersonation conditions). The group knew that they were dealing with images of people in disguise, and were informed of the true identity classification (same or different) in each matching situation to aid them in their decisions.

**Figure 5.7** Image taken during stimuli selection process.

In order to better understand what was driving the decisions made by the stimuli selection group I asked them the following questions for each identity in each disguise condition: 1). Which image provides the best match? 2). How good is the disguise (rated on a scale of 1 – 7 with 1 representing 'not a good disguise at all' and 7 indicating an extremely good disguise)? 3). What is it about the chosen image which makes it an effective disguise? The results of each of these are discussed in Chapter 6, Understanding Disguise, but for now I will continue to focus on the database itself and the image pairs that will be used to test face-matching performance.

**Face Image Pairs**

The images chosen by the stimuli selection group provided the final disguise image for each of the models in each of the disguise conditions. All other images were disregarded, with only the most convincing disguises for each model kept in the FAÇADE database to satisfy the disguise face image conditions. I was then able to create face image pairs for both disguise and no-disguise conditions using the database images. The face-matching task included the following face image pairing conditions: same identity no disguise, same

identity *evasion* disguise, different identity *similar* no disguise, different identity *similar impersonation* disguise, different identity *random* no disguise and different identity *random impersonation* disguise.


**Figure 5.8** Example pairs for each condition. Top row shows same identity pairs, the lower two rows show different identity pairs. Pairs in the first column are in no disguise. Pairs in the second column are in disguise. All 26 models were photographed in each of the conditions.

Examples of the stimuli pairs are shown in Figure 5.8. The image pairs on the left show no disguise image pairings. The top two rows in this column consist of pair types that are very similar to those seen in other standardised face-matching tasks such as the GFMT. Pairs of images can show either the same person or two different people. I have extended the design, through introducing the third column to include two types of different person trial. Similar pairing means that the foil here is the most similar looking person from the option of face images. Random pairing means that the foil is drawn at random from

images of models of the same sex. Therefore, random pairings can be rather dissimilar in appearance, showing two faces side by side, for which one would not normally expect the identities in the pair to be confused.

This led to a very convincing (and rather confusing) series of images within the FAÇADE database (see Figure 5.9).

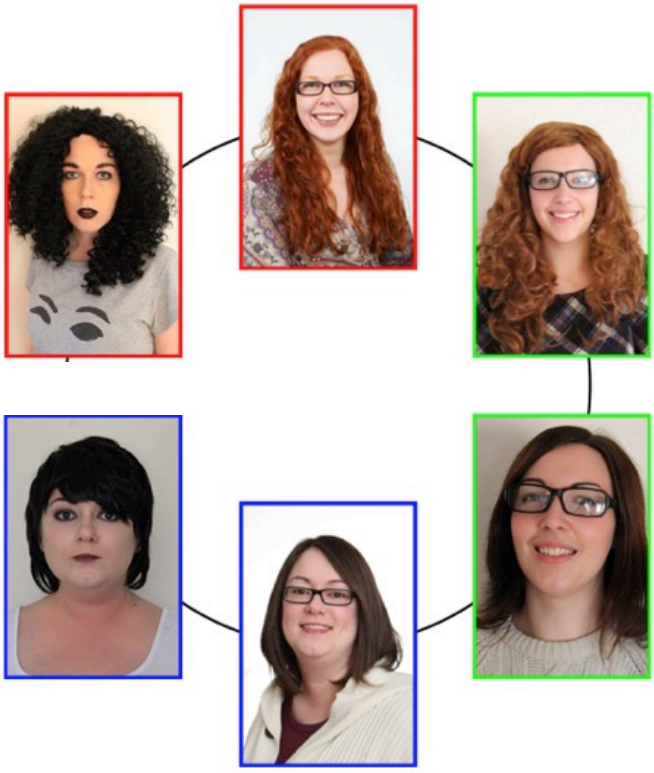

**Figure 5.9** Selection of images taken from the disguise base database to create a wheel of disguise. Images with the same colour frame show the same identity. Images with different colour frames are of different identities.

### Familiarity

Familiarity has been an overarching theme throughout this thesis. It has been repeatedly shown that people who are familiar with a face are better at matching images of that face than people who are unfamiliar with the face. Now the familiarity advantage is being pushed even further than before as I test whether familiarity with a face can help in the

case of deliberate disguise. In theoretical terms, as a face becomes familiar, the normally accepted range of variability for a face is being learnt. Effective disguise involves taking a face outwith that normal range of accepted appearance. It is thus possible that in the case of disguise, familiarity with the face or faces involved may not help much, precisely because disguising a face takes it out of the accepted range of appearances and hence out of the familiar viewer's area of expertise.

Dhamecha et al. (2014) argue that familiarity with a face does help with face-matching performance for disguised faces. In their experiment Dhamecha and colleagues (2014) define familiar viewers as participants who worked in the same department as the models whose images were used in the matching task, and unfamiliar participants as people who did not have previous encounters with the models. Although there is a significant effect of familiarity when the results of trials are pooled, the effect is not significant for both trial types considered in isolation. For same person trials, familiar viewers were more accurate at the task then unfamiliar viewers. However for different person trials, incorrect matches were made equally often by familiar and unfamiliar viewers. Both items in both match and mismatch pair items were most often disguised. Based on these findings, the familiarity advantage seems uncertain for the case of disguise.

I expect that unfamiliar viewers will make more mistakes on the disguised face trials than familiar viewers will, as their representation of these faces are less finely tuned than that of familiar viewers. Unfamiliar viewers frequently accept imposter faces to be a target face. This is demonstrated by both earlier work in this thesis (e.g. Chapter 2) and the results of face-matching tasks such as the GFMT, which show around 80-90% performance accuracy for unfamiliar viewers, confirming that people are making errors in identity judgment.

There are however instances of extremely familiar viewers being tricked by imposter scenarios. Thompson, (1986) demonstrated that it is possible to mistakenly reject a family member if context interferers with the identity judgment. Thompson, (1986) constructed a scenario where a daughter walked past her parents, without acknowledging them, while they were on holiday. The parents believed that their daughter was at home, in a different location. The parents concluded that the identity was not that of their daughter, based on the incorrect contextual information that they held. This example demonstrates that familiarity does not guarantee correct identity judgment. It is possible that disguising a face may have a similar effect, as some context may be lost when a face is disguised with the attempt of removing a face outwith its accepted range of appearances. For example, a person may normally be seen in work clothes with a professional and approachable appearance, but in disguise their displayed persona may be different.

**The Study – Research Questions & Predictions**

This study will use image pairs from the FAÇADE database to answer four main research questions. Firstly, I will address whether overall face-matching performance accuracy is worse for disguised than undisguised face pairs. Presumably it will be more difficult to match disguised faces than images of cooperative stimuli.

The second question of interest is whether all disguise manipulations cause equal impairment to face-matching performance or if disguise type has an effect. Results of past experiments show that more errors are made for same person trials than different person trials in face-matching tasks. This demonstrates difficulties in integrating multiple images of the same identity (e.g. Jenkins et al. 2011). Thus, a deliberate attempt to frustrate the integration of identity (the evasion condition) will likely make the task of 'telling the faces together' harder still. Therefore I predict that trials involving evasion disguise will be more error prone than impersonation disguise trials.

I will also investigate whether there is an effect of impersonation type, i.e. is it easier for someone to pass themselves off as being a person who looks naturally similar in appearance to them or are they equally effective at looking like somebody who is selected at random? Past research suggests that impersonation similar disguise will cause more errors than impersonation random disguise. Light, Kayra-Stuart & Hollander (1979) found that people were worse at remembering faces that looked similar to a prototype face than those that looked unusual in appearance, suggesting a role of distinctiveness. I thereby expect that impersonation random disguises will lead to fewer errors in the face-matching task than impersonation similar disguises.

Each of these questions will also be addressed with regard to familiarity with the faces viewed. As discussed in the familiarity section, based on theories of face learning, familiarity is not guaranteed to help in case of disguise. Familiar viewers will however be used to viewing the models' faces over a wider range of appearances than our unfamiliar viewers. Therefore familiar viewers are predicted to be better than unfamiliar viewers at the face-matching task, although familiar viewers may also be somewhat affected by the disguise manipulations. In total I will conduct three disguise face-matching experiments, each of which manipulate viewer group but will also address each of the research questions here outlined. Experiment 10 will test face-matching performance of unfamiliar viewers, Experiment 11 will test performance of unfamiliar viewers who are informed of the disguise manipulation and finally Experiment 12 will access face-matching performance for familiar viewers.

## 5.5 Experiment 10: Unfamiliar Viewers

Experiment 10 examines face-matching performance for unfamiliar viewers on a task involving disguised and undisguised image pairs. The experiment will use the images from the disguise face database, therefore allowing an investigation of face-matching performance accuracy for evasion, impersonation similar and impersonation random disguise.

I predict that disguise will impair face-matching performance, with evasion disguise making face pairs most difficult to correctly identify, and impersonation disguise will also cause difficulties but to a lesser degree. I believe impersonation similar disguises will be easier to execute than impersonation random disguises, hence more errors will be made when matching impersonation similar than impersonation random faces.

## Method

### Participants

30 undergraduate students (M = 8, mean age = 23) from the University of York volunteered as participants in this study in return for payment or course credits.

### Design & Stimuli

The experiment took the form of a 2x3 within subject design, with factors Disguise Condition (levels: Disguise and No Disguise) and Pair Type (levels: same, different similar and different random). The dependent variable was performance accuracy in the face-matching task for each of the independent variables listed above.

Stimuli were the face image pairs from the FAÇADE database. Stimuli were kept constant across all experiments in this chapter.

### Procedure

Participants viewed the image pairs from the FAÇADE face database on a computer screen. The viewing distance was 50cm from the screen. Image pairs were presented one at a time as part of a self paced face-matching task. The participants' task was to decide, for each image pair, whether the two images were of the same person's face or were

images of two different people – pressing ‘z’ on the keyboard for same pairs, and ‘m’ for different. All image pairs were presented in a randomized order and over the whole experiment the participants saw each of the models in each of the disguise conditions. In total participants viewed 156 image pairs.

It was important to check that all items in the task were unfamiliar to the participants in this experiment. Following the face-matching task, participants were given the reference images which they had viewed in the task, and asked to indicate how familiar they were with the person’s face before the experiment. Familiarity was measured using the familiarity scale that I designed in Chapter 1 of this thesis. Participants placed the face cards on the familiarity scale that ran across the desk, indicating familiarity on the scale from 0 (completely unfamiliar) to 100 (extremely familiar).

**Result**

Participants’ responses from the face-matching task were broken down into each of the Disguise Conditions and Pair Types, with a mean score of performance accuracy calculated for each. These means were then combined across all participants and averaged to calculate the overall mean for each condition. These breakdowns of scores by disguise condition and disguise type allowed the research questions of whether there was an overall effect of disguise on face recognition accuracy, and whether disguise type influence accuracy, to be answered.

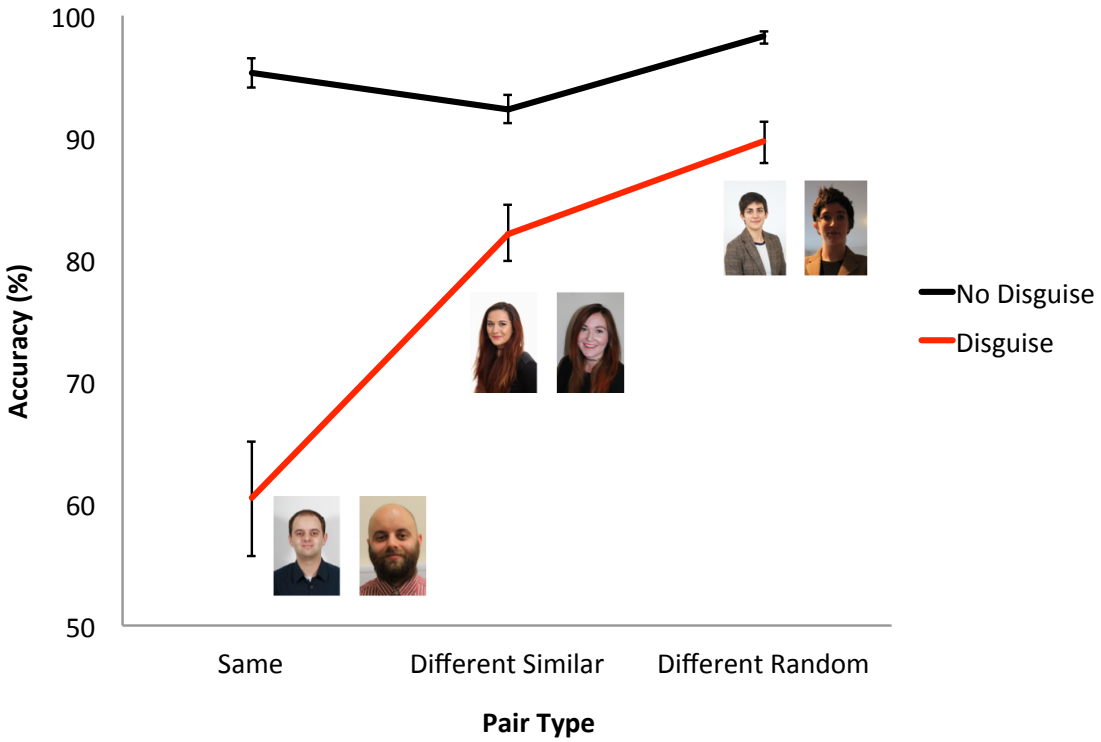


**Figure 5.10** Performance accuracy for unfamiliar viewers for evasion, impersonation similar and impersonation random pairs when the images consisted of no disguise or disguise pairs. Error bars show the standard error of the mean.

A within subjects ANOVA with factors of *Disguise Condition* and *Pair Type* was conducted to find out whether disguise presence affected face-matching performance and whether disguise *Pair Type* affected matching accuracy (see Figure 5.10). As predicted, there was a significant main effect of *Disguise Condition*. Participants performed more poorly for *Disguise* (M = 77.39 % accuracy, SD = 21.27) than *No Disguise* face pairs (M = 95.26% accuracy, SD = 5.97) [F(1, 29) = 75.88, p<.001, $\eta_p^2$ = .72]. There was also a significant main effect of Pair Type (F(2,29) = 22.87, p<.001, $\eta_p^2$ = .44] and a significant interaction between *Disguise Condition* and *Pair Type* [F(2,29) = 37.95, p<.001, $\eta_p^2$ =,57]. This showed that both the presence of disguise and also the type of trial affected performance accuracy rates, but accuracy was not influenced in the same way by *Pair Type* (*Same, Different similar,* or *Different random)* in each of the disguise conditions.

Simple main effects were conducted to find out where the significant difference lay. *Pair Type* was significant for *Disguised* [F(2,116) = 53.79, p<.001, $\eta_p^2$ = .48] but not for *No Disguise* pairs [F(2,116) = 2.03, p>.05, $\eta_p^2$ =.03], demonstrating that participants showed a difference in their accuracy of matching across different pair types only for the disguise present faces. Disguise presence significantly affected performance for each of the *Pair Types*; *Same* identity pairs [F(1,87) = 151.5, p<.001, $\eta_p^2$ = .64], *Different* (*Similar* appearance) identity pairs [F(1,87) = 12.78, p<.005, $\eta_p^2$ = .13] and *Different* (*Randomly* matched) pairs [F(1,87)= 9.19, p<.005, $\eta_p^2$ = .1].

Tukey tests revealed that there were significant differences between each level of trial type for the disguise present faces. *Same* identity pairs (represented on the graph as *Same*, condition *Disguise*) caused most errors (M = 60.38% accuracy, SD = 25.69), then *Different Similar* identity pairings (those on the graph shown as Different, and condition Disguise) (M = 82.18% accuracy, SD = 12.8), and finally *Different Random* identity pairings (M = 89.62% accuracy, SD = 9.32).

**Discussion**

This experiment found that disguise did significantly impair face-matching performance for unfamiliar viewers. Therefore, in relation to my first research question, and in accordance with the hypothesis, matching disguised faces is more difficult than matching faces that are not in disguise. My second research question, which asked whether all disguise manipulations cause equal impairments to face recognition, is also addressed by this experiment. The results were again as predicted, with evasion disguise affecting face-matching accuracy more than impersonation disguise. With regards to research question three, there was also a significant effect within impersonation itself. Unfamiliar viewers made more matching errors when the Impersonations involved someone who was of a similar appearance to the model than when the impersonation pairings were at random.

The results from this experiment will later help to answer research question four, regarding matching accuracy and familiarity, providing data from unfamiliar viewers which can be compared with that of familiar viewers when collected.

These findings using the FAÇADE database address the issue of disguise face-matching in more detail than any previous study. However, I want to continue to understand disguise even further. I found that face-matching performance was poor for disguised faces compared to faces that were not in disguise, and although difficult to compare directly with other studies, performance for the disguised faces in my task was poorer than performance on the full version of the GFMT - mean performance accuracy was 77% for the disguised face pairs, whereas performance on the full version of the GFMT is 87%. Performance may have been particularly poor in my experiment because viewers were unaware that faces they viewed may have been disguised. It is possible that knowledge of the possibility of disguise presence could improve face-matching performance.

## 5.6 Experiment 11: Unfamiliar (Informed) Viewers

In Experiment 10 unfamiliar participants' face-matching performance was impaired by disguise presence. Making accurate identity matching judgments for the disguised faces was a difficult task, but it is possible that this task was difficult because participants were not expecting to see faces that were disguised. Expectations have been reported to influence performance accuracy in visual search and other cognitive tasks (e.g Alain & Proteau, 1980; Kastner, Pinsk, De Weerd, Desimone & Ungerleider, 1999). If participants are told about the disguise aspect of the study then performance may improve. This will now be tested in Experiment 12.

If poor performance from Experiment 10 is because participants did not know to look for disguises, their performance accuracy should be far better in Experiment 11 than it was in Experiment 10. If performance accuracy for disguised faces is no better when participants

have this knowledge, then being naïve to the disguise component is not what makes the task difficult.

The research questions remain the same as in Experiment 10, but this time I investigate whether the results of Experiment 10 remain when participants are informed of disguise.

## Method

### Participants

Thirty undergraduate students from the University of York who had not taken part in previous experiments involving the FAÇADE database, volunteered as participants for this experiment (mean age = 21, M = 11).

### Design & Stimuli

The experiment took the form of a 2x3 within-subject design, with factors Disguise Condition (levels: Disguise and No Disguise) and Pair Type (levels: Same, Different Similar and Different Random). The dependent variable was performance accuracy in the face-matching task for each of the independent variables listed above. This will enable the research questions outlined in the introduction to be addressed.

To test whether performance differs for disguised faces between Experiment 10 and Experiment 11, a between subjects design will be used.

Stimuli were the face image pairs from the FAÇADE database. Stimuli are kept constant across all experiments in this chapter.

**Procedure**

Experiment 11 was carried out exactly like Experiment 10, the only difference to the method being that this time participants were made aware before beginning the experiment that some of the face images they view may be disguised to either look unlike themselves or to look like another person. It was clearly stated that the face-matching decision always concerned an identity judgment based on the true identity of the face. Participants were to decide if the images they saw were really images of the same person, or if they were images of two different people.

**Results**

Results were cacluated exactly as in Experiment 10, with mean performance accuracy broken down by Disguise Condtion and Pair Type. In order to answer the research questions regarding an overall effect of disguise (Disguise Condition), possible differences between Pair Type, a within subjects ANOVA was conducted.

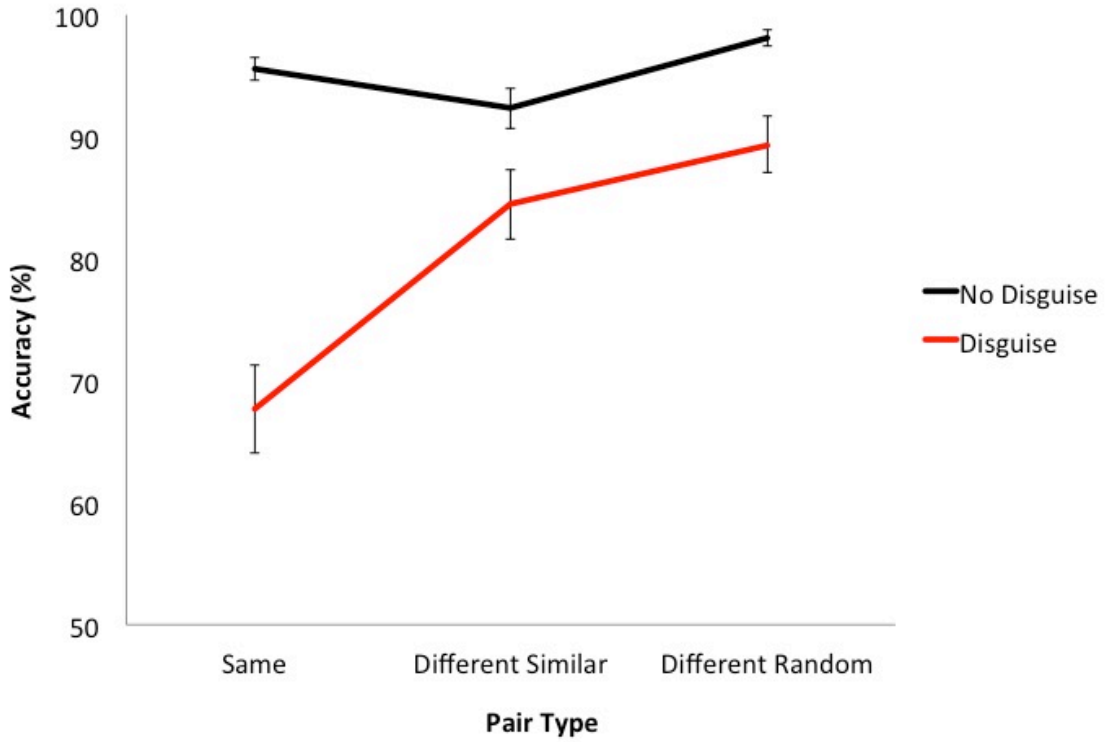


**Figure 5.11** Performance accuracy of unfamiliar viewers who were aware of the disguise component of the face-matching task. Error bars show the standard error of the mean.

A significant main effect was found for *Disguise Condition* - participants were worse at matching *Disguised* faces (M = 80.51, SD = 18.68) than faces with *No Disguise* (M = 95.30% accuracy, SD = 6.62) [F(1,29)= 44.25, p<.001, $\eta_p^2$ = .6]. There was also a significant difference main effect of *Pair Type* [F(2,29)= 21.39, <.001, $\eta_p^2$ = .42] and a significant interaction between *Disguise Condition* and *Pair Type* [F(2,29) = 55.52, <.001, $\eta_p^2$ = .66] (see Figure 5.11).

Simple main effects highlighted there was a significant effect of *Disguise Condition* on performance for each of the Pair Types (same [F(1,87) = 119.56, p<.001, $\eta_p^2$ = .58], *Different Similar* [F(1,87) = 9.45, p<.005, $\eta_p^2$ =.1] and *Different Random* [F(1,87) = 11.74, p<.001, $\eta_p^2$ = .12]). There was also a significant main effect of *Pair Type* for both *Disguised*

[F(2,116) = 56.21, p<.001, $\eta_p^2$ = .49] and *No Disguise* faces [F(2,116) = 3.64, p<.01, $\eta_p^2$ = .06].

Tukey tests revealed that there was a significant difference in performance between each level of Pair Type for the Disguise faces. Same identity pairs (represented on the graph as Same, condition Disguise) caused most errors (M = 67.60% accuracy, SD = 20.01), followed by *Different Similar* identity pairings (those on the graph shown as Different, and condition Disguise) (M = 84.49% accuracy, SD = 15.73), and finally *Different Random* identity pairings (M = 89.36% accuracy, SD = 10.09). For *No Disguise* faces the only significant differences lay between performance accuracy for *Different Similar* (M = 92.31% accuracy, SD = 8.86) and *Different Random* face pairs (M = 98.08% accuracy, SD = 3.46)

In order to test whether performance is better if participants know of the possibility of disguise manipulations than when they are naïve to the disguise component, I carried out a between subjects ANOVA to compare performance accuracy for disguised faces in Experiment 10 with that of Experiment 11. Interestingly participants in Experiment 11, who knew that some of the faces in the experiment were disguised to not look like themselves or to look like a specific other person, were no better at the face-matching task with this knowledge of the disguises, than participants in Experiment 10 who were not informed of the disguise manipulations before the experiment F(1,58) = .625, p = .43, $\eta_p^2$ =.01. Thus, knowing to look out for disguise does not help to see through disguise.

**Discussion**

The same pattern of results was seen for each of the research questions for Experiment 11 (when participants were aware of the disguise element) as for Experiment 10 (when participants were not aware that faces could be in disguise). Experiment 11 found that informed viewers were significantly worse at matching disguise face pairs than faces that

were not in disguise, evasion caused more difficulty than impersonation disguise, and impersonation similar trials were harder than impersonation random pairs.

The unique research question for Experiment 11 was whether performance accuracy for disguised faces would be better than it was in Experiment 10. No significant difference was found in performance for matching disguised face pairs in each of these experiments, therefore knowledge of disguise does not make the task easier. Alternatively, it is possible that participants in Experiment 10 may have figured out the disguise manipulation for themselves whilst doing the task, either way performance remains poor for disguised faces in both the experiments, suggesting that matching disguised faces is an extremely challenging task regardless of holding knowledge that disguises may be present amongst the image pairs.

## 5.7 Experiment 12: Familiar Viewers

There is a huge distinction in the face recognition literature between face-matching performance for familiar, compared with unfamiliar viewers. Familiar viewers perform with impressive levels of accuracy in matching tasks even when the images available for comparison are of poor quality (Burton et al. 1999). Chapters 2 and 4 show further situations where familiarity improved performance in challenging face-matching tasks. I now test familiarity on perhaps our most challenging task to date – the situation of deliberately disguised faces. Can familiarity with a face help us see through deliberate disguise and hence make the viewer immune to deception of disguise?

Jenkins et al. (2011) highlight that familiar viewers are able to easily see through natural changes in a face. The distinction between familiar and unfamiliar face recognition performance may be due to experience of variability in photos of the same face. They argue that when someone is becoming familiar with a face they are learning all of the different appearances, which that face can take. Any face that falls within the excepted

range of faces for a specific person will be considered a match for that identity, any image which does not fall within the range of accepted appearances for a face will not be considered a match. It is believed that the range of accepted faces for a given face is smaller for unfamiliar viewers and larger for familiar viewers who have had more experience with the face. Therefore the idea is that the acceptable range of facial appearances for a given person becomes more finely tuned as their face is learned. This explains why familiar viewers can see through changes in viewpoint, expression and lighting, whereby unfamiliar viewers can be impaired by trivial image changes - they do not have enough experience with the face to accurately know what appearance the face would take under these changes (Bruce 1982; Adini, Mosses & Ullman, 1997). However, this theory does not necessarily cover the situation of disguise.

Presuming that the viewer has not had prior exposure to an individual in their disguised form, it is unknown whether the previous exposure they have had with the face would be sufficient to also help the viewer see through a disguised version of the face. Essentially this is the challenge for creating a convincing disguise. For an evasion disguise to fool a familiar viewer it must change the appearance of the model so that the new appearance falls out with the accepted range of appearances held by the viewer for that model. Impersonation disguises must move outwith the accepted range of the model and into the accepted range of the reference photograph in order to cause a familiar viewer to make face-matching errors. If the models have been effective in doing this, familiarity with a face may not help a great deal, as the face would be taken outwith the familiar viewers area of expertise.

There are essentially two strands of research questions regarding familiarity. The first includes the familiar participants alone, and asks whether familiar viewers are able to see through disguise manipulations. This relates to the first three research questions laid out in the introduction. If familiarity allows participants to see through disguise, then there would be no difference in performance accuracy for disguised and no disguise face pairs. If however, familiar viewers are worse at matching disguised face pairs than undisguised

face pairs, then even familiar viewers are affected by an overall disguise manipulation. It is again possible that some types of disguise may affect performance more than others. For familiar viewers I predict that impersonation disguise will not be as effective as evasion disguise. To identify faces in impersonation disguises, participants can use both of the faces in the pair to look for identity cues, as they are familiar with both of the faces. Therefore a model's task is harder as they have to look both unlike themselves and convincingly like the person they are trying to impersonate, whereas evasion involves moving from only the accepted range for a person's own appearance, which can be achieved in many different ways.

The second strand will answer research question number four, of whether familiarity with a face aids disguised face-matching. This will involve a comparison of results for disguised face-matching performance across all of the Experiments conducted in this chapter, thus comparing familiar viewers with unfamiliar viewers. As familiar viewers have more experience than unfamiliar viewers with the faces in the FAÇADE database, I predict that participants will be significantly better than the unfamiliar viewers at matching disguise faces, simply because they have seen the faces in a wider range of appearances, and therefore may be better able to see through some aspects of the disguises.

## Methods

### Participants

The participants in this study were 30 individuals who were colleagues and friends of the disguise models (M = 14, mean age = 27). Participants saw the models on almost a daily basis; therefore they were of high levels of familiarity with the models' faces.

**Design & Stimuli**

The experiment took the form of a 2x3 within subjects design, with factors *Disguise Condition* (levels: *Disguise* and *No Disguise*) and *Pair Type* (levels: *Same*, *Different Similar* and *Different Random)*. The dependent variable was performance accuracy in the face-matching task for each of the independent variables listed above. This will enable the research questions outlined in the introduction to be addressed.

To test whether face-matching performance for disguised faces differs as a result of familiarity, a between experiment analysis will be conducted comparing the results of Experiment 10 and Experiment 11, with those of Experiment 12.

**Procedure**

Procedure was as in Experiment 10.

**Results**

A 2x3 ANOVA was conducted to investigate the effect of Disguise Condition and Pair Type on familiar participants' face-matching performance.

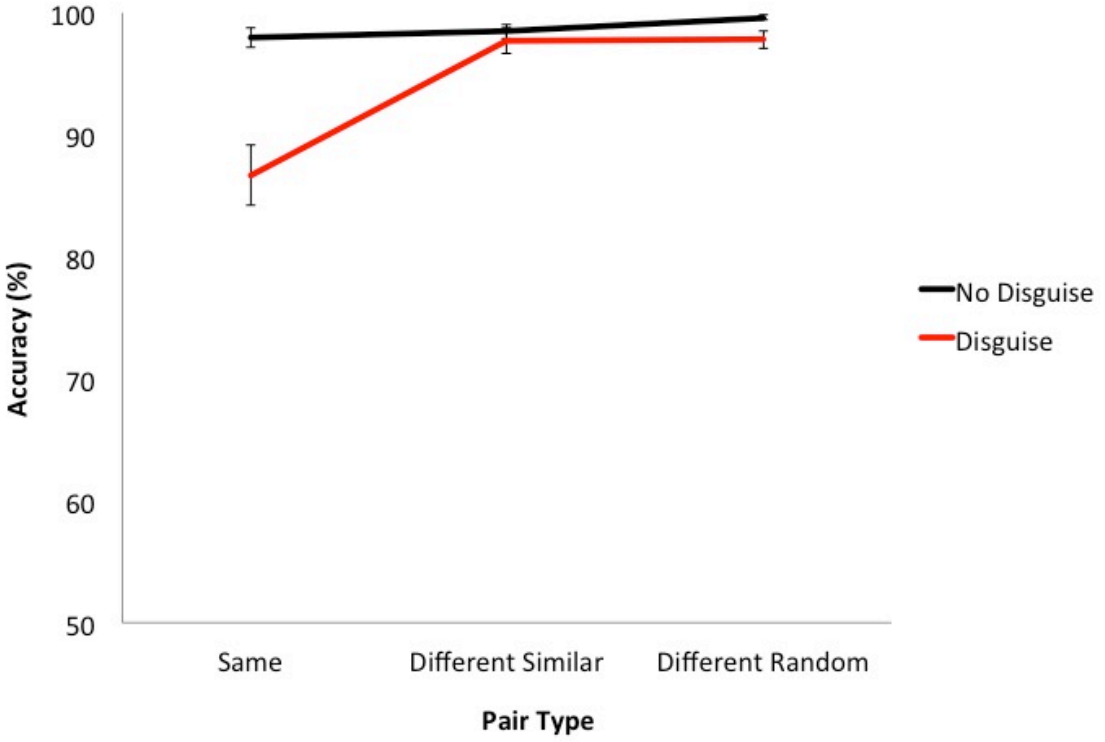


**Figure 5.12** Performance accuracy in the face-matching task for viewers who were familiar with the models whose images featured in the task. Error bars show the standard error of the mean.

A significant main effect of both *Disguise Condition* (No Disguise or Disguise) [F(2,58) = 20.01, p <.001, $\eta_p^2$ = .41] and *Pair Type* (*Same*, *Different Similar* and *Different Random*) [F(1,29) = 24.99, p <.001, $\eta_p^2$ = .46] were observed. This shows that familiar viewers were worse at matching disguised faces than faces that were not in disguise, and the type of disguise made a difference to performance accuracy. There was also a significant interaction between *Disguise Condition* and *Pair Type* [F(2, 58) = 22.23, p <.001, $\eta_p^2$ = .43], meaning that participants were not equally affected by disguise presence for all disguise pair types (see Figure 5.12).

To understand where disguise caused impairments to face-matching performance simple main effects were calculated. These revealed that the only significant differences for *Pair*

*Type* were for the disguised faces [F (2,116) = 40.97, p<.001, $\eta_p^2$ = .41), and not for *No Disguise* faces, p>.05. Specifically, Tukey tests revealed that familiar viewers were significantly worse at matching disguised evasion (*Same identity*) faces than impersonation (*Different* identity) pairs (M = 86.67% accuracy, SD = 13.54), with no significant difference in performance levels for matching *Impersonation Similar* (M = 97.69% accuracy, SD = 5.59) compared with *Impersonation Random* pairs (M = 97.82% accuracy, SD = 4).

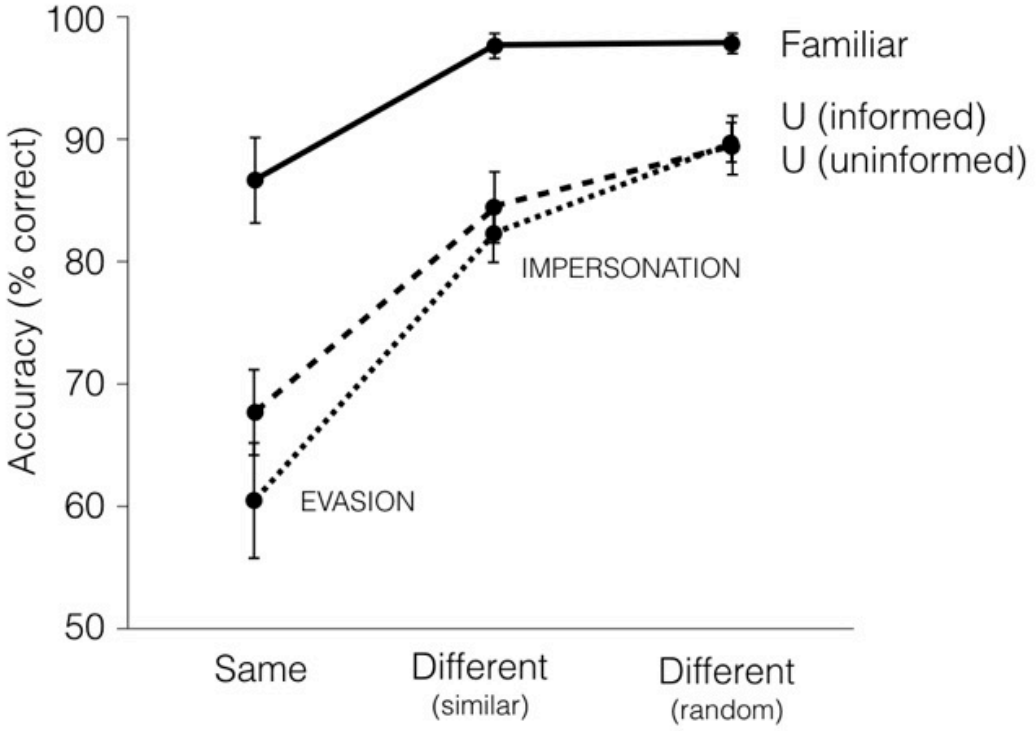


**Figure 5.13** Graph showing performance accuracy for Disguise face pairs for each of the 3 Experiments: U informed (Experiment 10), U uninformed (Experiment 11), Familiar (Experiment 12).

The effect of *Familiarity* on performance accuracy for disguised faces can be examined by comparing results across the three experiments (Experiment 10 - Unfamiliar, Experiment 11 - Unfamiliar informed and Experiment 12 - Familiar) (see figure 5.12). Comparing performance of all experiments on disguised face pair types showed a strong significant main effect of Experiment, F(2, 87) = 19.15, p<.001, $\eta_p^2$ =.31.

Pairwise comparisons revealed that *Familiar* viewers were significantly better at matching disguised faces than both *Unfamiliar* viewers (Experiment 10) and *Unfamiliar Informed* viewers (Experiment 11) and reiterate no difference in performance accuracy in matching disguised faces between *Unfamiliar* and *Unfamiliar Informed* viewers. The mean difference in performance accuracy between *Familiar* viewers and *Unfamiliar* viewers (Experiment 10) was 17%, CI = 10.8 – 22.53, p<.001. And between the *Familiar* viewers and *Unfamiliar Informed* viewers (Experiment 11) this was 13.5%, CI = 7.68 – 19.41, p<.001. Mean difference in performance between the two unfamiliar groups on the disguise present matching trials was 3%, this was not a significant difference, p>.05.

**Familiarity Score Comparisons**

In each of the experiments in this chapter, participants indicated their familiarity with each of the faces included in the face-matching task after they had completed the face-matching task. Scores were based on familiarity prior to the experiment. I compared mean familiarity score with the face items, for each participant, across the three experiments using a one-way between subjects ANOVA. The ANOVA showed a significant difference in familiarity between the groups [F(2,87) =314.43, p<.001, CI = 22.06 – 37.08. Post hoc tests revealed that the familiar viewers were significantly more familiar with the faces (M = 75.56 familiarity rating) than both the unfamiliar viewers (M = 6.81 familiarity rating, mean difference = 68.75, SE = 3.19, CI = 60.94 – 76.55, p<.001) and the unfamiliar informed viewers (M = 4.69 familiarity rating, mean difference = 70.87, SE = 3.25, CI = 62.92 – 78.81, p<.001). There was no significant difference in the familiarity ratings for the two unfamiliar groups (mean difference = 2.12, SE = 3.25, CI = -5.82 – 10.06, p>.05).

**Discussion**

With reference to research question 1, which posed the question of whether disguise impairs face-matching performance, familiar viewers were significantly worse at matching

disguised face pairs than faces that were not in disguise. Whereas previous face-matching tasks find ceiling performance levels for familiar viewers, in the case of disguised faces, familiarity can not completely overturn the effect of disguise. Research question 2, whether all disguise types are equally challenging, is particularly relevant in the case of familiar viewers. The familiar viewers were worse for disguised faces, only when these were evasion faces, with no significant difference in performance between disguise and no disguise recognition for impersonation faces. Question 3 considers differences between types of impersonation disguise. Unlike in the previous experiments, impersonation pair type did not matter – familiar viewers could see through both types of impersonation disguise equally well.

Research question 4, searching for an effect of familiarity, was also answered by this experiment, in the between experiment analysis. Familiarity improved face-matching performance for disguised faces compared to that of unfamiliar viewers. Familiar participants (Experiment 12) were 15% better than unfamiliar participants (averaged result of Experiment 10 & Experiment 11) at correctly identifying same person and different person disguise image pairs. Familiar viewers were 11% better than the average result of the two unfamiliar viewer experiments at matching impersonation faces and 23% better at matching evasion face. These results suggest that when identity judgments involving disguise need to be made, recruiting familiar viewers to make the identity judgment would likely lead to a more accurate judgment of identity.

These findings fit with my theoretical interpretations of both familiarity and effective disguise. It has been argued that when people become familiar with a face what they are learning is all the different ways that that face can look (Jenkins et al., 2007). For familiar viewers then, they will be familiar with a far greater range of appearances that any one of our models' face can take, than the unfamiliar viewer group will be. This may mean that some of the changes to appearance, perhaps expression and pose changes for instance, are learnt appearances of the face in question for the familiar viewers, and hence would not disguise the face as it may for unfamiliar viewers. Previous exposure of familiar

viewers to some of the facial changes used by the models to achieve their disguises, likely explains some of the difference in performance between familiar and unfamiliar viewer groups. Any experience that a familiar viewer had with any of these model's appearance changes would be have already been stored within the accepted range of appearances for that model's face. Thus, any variation, which a familiar viewer had been previously exposed to for the faces in disguise, would not act as an effective disguise manipulation for a familiar viewer, but it would for an unfamiliar viewer.

Although familiar viewers were somewhat impaired by evasion disguise, they could easily see through imposter disguise. For an imposter disguise to be effective the model must have moved both outwith their own face and into the face space of the person they are trying to impersonate. Direction of disguise is limited to moving only towards (and into) the face space of the other person, so disguise is automatically more difficult than in the case of evasion whereby appearance can be changed in any direction that brings the face image outwith its normally accepted face space. The familiar viewers were familiar with both the imposter and target person who featured in each trial. This allowed viewers to approach the task from two angles – the imposter face could be encoded to be the real person behind the imposter, or it could be believed to be not a good enough match to the target face to convincingly fall within their face space. Unfamiliar viewers were limited to using only the second of these strategies. These factors help to explain why familiar viewers perform with higher accuracy than unfamiliar viewers for both evasion and impersonation disguise and also highlight reasons for evasion disguise causing familiar viewers difficulty whereas impersonation disguise did not.

## 5.8 General Discussion

This chapter saw the creation of the FAÇADE database, which consists of images of male and female human faces in disguise (evasion and impersonation) and no disguise image conditions. Disguises were all model led, meaning that the models were free to disguise themselves as they wished and disguises were not limited to the use of props. Image pairs

from the FAÇADE database were then used to test face-matching performance accuracy for both familiar and unfamiliar viewers. In Experiments 10, 11 and 12 deliberate disguise had an effect on overall face-matching performance accuracy – participants made more identity judgment errors on disguise pairs than no disguise image pairs. Disguise pair type also affected performance accuracy. Evasion disguises resulted in more errors than impersonation disguise, and for unfamiliar viewers only (Experiments 10 & 11), impersonation similar trials were more error prone than impersonation random trials. These results show that disguise is more convincing when looking *unlike* yourself than when looking *like* a specific other person. For the case of unfamiliar viewers, impersonation disguises are more effective if the target looks similar to the model prior to disguise. In Experiment 2, I found that unfamiliar viewers were no better at matching disguised faces when they had been informed before completing the face-matching task that some of the images would involve disguise, than they were when they were naïve to the disguise element of the study (Experiment 10). In Experiment 12, a familiarity advantage was found. Familiar viewers were better at correctly identifying match or mismatch disguise face pairs than unfamiliar viewers (Experiments 10 & 11). Familiar viewers easily saw through impersonation disguises, but even for familiar viewers, evasion disguise caused significant impairment to face-matching performance.

Face matching performance for disguised face images was much lower than performance from the GFMT, which is known as the standardised test of face-matching performance. Viewers performed with around 89% accuracy on the long version of the GFMT and 83% on the short version whereas here mean performance for all disguised faces was 77% for unfamiliar (uninformed) and 80% for unfamiliar informed viewers. Furthermore, familiarity has led to ceiling levels of performance in past face-matching tasks (e.g. Burton et al. 1999), however performance of familiar viewers for evasion disguise was even lower than that of unfamiliar viewers for cooperative stimuli in the GFMT. These numerical comparisons highlight how severely disguise impairs face-matching performance, with evasion disguise impairing even familiar viewers.

This study overcame the main limitations of past disguise research through use and creation of the FAÇADE database and also the experimental design. For example, Patterson and Baddeley (1997) Experiment 1, tested face memory for disguised images against exact image memory performance. I test performance based on a disguise versus no disguise face-matching accuracy comparison. The images included in the no disguise condition were not the same image of a face. Additionally, the images used in Patterson & Baddeley's Experiment 1 disguise condition could have shown two different disguises for the same identity – one at learning and a different disguise at test. I always compare a disguise to no disguise image matching comparison rather than introducing two different disguises. I believe having only one of the comparison images in disguise is better suited to real world application.

The overarching result of this chapter is that disguise presence impaired face-matching performance accuracy. This finding is in line with all previous research on disguise that was discussed in the introduction of this chapter. Disguise had previously been shown to impair face memory performance (Patterson & Baddeley, 1977; Terry, 1994; Righi et al. 2012). Dhamecha et al. (2014) also reported that face-matching accuracy for disguised faces was lower than accuracy levels reported for the GFMT, which is a standard test of cooperative face-matching. Dhamecha et al. (2014) tested matching for image pairs whereby both of the images in the pair were generally disguised. My studies show an effect of disguise on face-matching when just one of the images in the pair was in disguise, showing how difficult disguise can make the task. Additionally, the stimuli used by Dhamecha and colleagues (2014) included occlusion of features. Occlusion has been a common feature of past disguise databases, and it seems obvious that face-matching performance would be hampered when parts of the face are obscured. Testing face-matching performance using the FAÇADE database showed that that even when disguises do not occlude facial features, disguise makes matching faces a more difficult task than when the same identities are presented undisguised.

This chapter advances upon previous disguise research through the investigation of impersonation as well as evasion disguise. All past research discussed has looked at only the case of evasion disguise (Patterson & Baddeley, 1977; Terry, 1994; Righi et al. 2012; Dhamecha et al. 2014). I show that evasion and impersonation disguise both cause matching difficulties, but evasion and impersonation do not cause equal levels of difficulty. I was able to break these results down further as my design incorporated two types of impersonation – similar and random. Unfamiliar viewers were better at matching random disguise pairings than pairings where the model looked naturally similar to the target they impersonated. Familiar viewers could see through both types of impersonation. These findings highlight the importance of investigating impersonation disguise and evasion disguise. The different results for each confirm that disguise cannot be treated as one.

The findings from this chapter are also in agreement with previous face variability studies. Past research has shown that integrating multiple images of the same identity ('telling people together') is a difficult task (Jenkins et al. 2011). I have demonstrated that people are worse at matching evasion face images (telling people together) than impersonation faces (telling people together). As people are poor at integrating multiple images of a face when there is not a deliberate attempt to make the identity look different across photographs, it is not surprising then that performance is even poorer for same person trials when there is a deliberate attempt to evade identity.

A difference however between the previous findings on face variability and the findings of this chapter, relates to the effect of familiarity. Jenkins and colleagues' (2011) familiar viewers could easily group together multiple images of the same identity. In my face-matching task familiarity did not completely overcome the difficulty of matching disguised faces. Participants were significantly worse at matching face pairs where the model evaded their own identity than they were at matching faces where there was no disguise. Levels of performance for evasion disguise face pairs were also far lower than in previous studies that have tested cooperative face-matching performance of familiar

viewers (Burton et al. 1999). As familiarity did not eliminate errors for evasion in my face-matching task, this finding suggests that disguise took the faces outwith the area of expertise that familiar viewers held for the faces involved.

It would have been interesting to test for graded effects of familiarity in this Chapter, much like in Chapter 1. However, due to the participant samples that we used, familiarity was skewed towards the extreme ends of our familiarity spectrums for each of our experiments. Over 80% of the unfamiliar viewers rated the face items as completely unfamiliar, whereas over 80% of our familiar viewers were of very or extreme familiarity with the faces in the task. In addition to this, the scales were likely not directly comparable as familiar viewers were more conservative on their familiarity ratings than unfamiliar viewers due to their baseline level of familiarity with the faces. All familiar viewers were at least somewhat familiar with all of the images as the models were of their colleagues. This meant that those faces that qualified as extremely familiar were usually who the viewers had an exceptionally high level of interaction with (e.g. in the same office with, close friends with) where as people ranked lower down on the scale would also be very well known, and likely classified as extremely familiar in many scenarios, but were less familiar in relation to the other faces in the task. It is therefore possible that even if a graded analysis was attempted on the small numbers in the lower familiarity bands, that the graded result may not be found in this experiment given that all familiar viewers were of a baseline high familiarity with the faces concerned.

This thesis has focused exclusively on conducting studies of face-matching performance for various challenging situations. Face-matching is a current and common security scenario, thus warranting thorough investigation, however it would also be interesting to assess the situation of face memory for disguised and undisguised faces. Past studies attempted to address this question but used image matching comparison scenarios and questionable disguise stimuli. It would be interesting to find out whether face memory is worse for disguised faces in our database. The case of impersonation disguise would be particularly thought provoking as this has not received any past investigation and it is

unclear whether participants would be confused by seeing the target face or model in undisguised form and be able to make links to the impersonator from the impersonation disguise image or if they would incorrectly identify the target face as the face seen before.

Future research could test human versus machine performance for the FAÇADE database pairs. Previous research has tested machine performance for recognising disguised faces, generally testing the performance of several different algorithms against each other (e.g. Singh et al., 2009). Dhamecha and colleagues (2014) showed that machines performed at a level similar to the unfamiliar human viewers for their face pairs that included faces in disguise, but looked at only the case of evasion disguise and for an unrealistic set of disguise faces. My database would allow a direct comparison between unfamiliar and familiar human viewer face-matching performance with performance of computer algorithms for evasion, impersonation similar and impersonation random faces. Human performance has been successfully fused with algorithms in past studies (O'Toole et al. 2007) therefore if machine performance was tested on the FAÇADE database, fusing performance from machines and humans may be a method which could be used to improve disguise face recognition performance also.

Implications of this research are that care needs to be taken over disguise faces and methods need to be explored to find out ways to improve disguise face recognition performance. It would be interesting to test the performance of super-recognisers on the task as super-recognisers have been found to outperform controls on other difficult tasks of face-matching (Robertson, Noyes, Jenkins & Burton, 2016). Disguised face recognition performance is poorer than performance for faces that are not disguised, hence more errors are likely to be made in real life deliberate disguise fraud situations than if someone is trying to pass themselves off to be someone else but has not made any deliberate attempt to change their appearance to look more alike. Poor performance for evasion disguise faces does however provide hope for undercover police investigations. Police Scotland had addressed concerns about their undercover police officers' identity

being recognised, but our findings suggest that evasion disguise is effective especially for unfamiliar viewers. However the converse of this finding is that criminals who evade their own identity provide a difficult face recognition scenario for the police.

I created a disguise face database, which is the first to include free evasion, and impersonation disguises on real female and male faces. This has allowed, and will continue to allow, disguise face recognition to be explored in more detail than ever before. Disguise impaired face recognition performance, with the type of disguise (evasion, impersonation similar or impersonation random) affecting face-matching accuracy. Familiar viewers outperformed unfamiliar viewers at the disguise matching task, however even familiar viewers were significantly affected by the presence of evasion disguise. Implications of these findings are that familiar viewers will likely be more accurate at tasks involving identity judgement than unfamiliar viewers. Poor disguise face-matching is advantageous in terms of undercover policing but also poses serious security threats if either evasion or impersonation disguises are used in criminal situations.

Performance accuracy on the face-matching task has now been established, but it is not yet known how people disguise themselves and which disguise manipulations work. Disguised face images in the FAÇADE database were created from photographing models posed in evasion, impersonation similar and impersonation random image conditions. Importantly for these new research questions, models were free in their disguise. This means that it was the models themselves who decided upon how to create each of the disguises and props were provided and arranged on request of the models rather than the experimenter. This free disguise element is unique to my database and allows an investigation of what changes people naturally make to disguise themselves. Additionally, to try and uncover the steps necessary to create an effective disguise, disguise face-matching data from this chapter can be combined with written reports from both the models and viewers regarding disguise manipulations and effectiveness. These questions will be explored in Chapter 6.

# Chapter 6 – Understanding Disguise

## 6.1 Chapter Summary

In order to further explore disguise - specifically how people disguised themselves, whether effective disguises could be identified, and which disguises worked - data from the matching task (Chapter 5) was analysed along with models' records of how they created their disguises and effectiveness ratings from an unfamiliar viewer group.

This chapter points to clear distinctions between evasion and impersonation disguise – both in terms of what changes people make, and also what makes a disguise effective. Unfamiliar viewers could accurately predict disguise effectiveness from viewing the target and reference image side by side.

Previous studies on disguise investigated the effect of adding props to a face, which often occluded facial features. I found that when models were free to create their own disguises they used far more methods other than simply adding props to do this. Evasion disguise revolved around creating differences with a target face in terms of internal and external features, through the use of makeup, clothing and hairstyle change sometime through wigs, and also using techniques such as expression and lighting change. Successful evasion disguise was also linked to creating differences in social inferences. Impersonation disguise on the other hand involved creating and focusing on similarities with a target face, but these similarities were related to physical changes to a face, for example internal and external features. Social inferences did not change to match those of the person being impersonated. It is evident that free disguise creation, especially in the case of effective disguise, is far more complex than purely the addition of props. It is important that disguise research acknowledges distinctions between evasion and impersonation disguise.

## 6.2 Introduction

As explained and demonstrated in Chapter 5, I have created a disguise face database consisting of 26 models, disguised to evade their own identity and to impersonate the identity of two reference individuals – one of naturally similar appearance to the model and the other selected at random, from images of other models of the same gender. These models were highly motivated, aided by the incentive of a performance based cash reward, to create extremely convincing disguises. Models were free to disguise themselves as they wished and could request props to assist their efforts. I provided no guidance on how the models should disguise themselves, although I did alert them to the different manipulations they could make using camera angle and distance (e.g. findings of Chapter 4). The only limitations on disguise were that the end result must look like a realistic I.D. photograph rather than a person in a fancy dress costume and that any props that would have to be removed in a passport security check were disallowed (e.g. hats and sunglasses). As demonstrated in Chapter 5 this led to the creation of a much more sophisticated disguise database than previous studies have used. It is of interest to better understand disguise - what the models did to disguise themselves in each condition, which disguises were believed to be the best and what actually makes for the best disguise (with reference to items that caused most difficulty in the matching task). I have matching performance accuracy rates for the disguise face database (Chapter 5). I also have information on what the models did to create their disguise and ratings from an independent viewer group for the effectiveness of each disguise. This data can be used to help to gain the desired better understanding of disguise.

As reviewed in Chapter 5, past research on disguise is limited in that it has not looked at impersonation disguise, and models whose images provided the disguise stimuli used simple disguises, adding props to occlude features. Perhaps the closest to background literature as a starting point to predicting what may make the best disguise, is to look at what image changes are already known to affect face recognition performance. Previous memory studies focus on the case of same person face recognition, therefore apply more

to evasion disguise rather than impersonation. Most previous studies have focused on the effect of changing expression, viewpoint or lighting on face recognition performance.

**Factors that Increase the Difficulty of Face Recognition**

The literature paints a clear picture that face recognition performance accuracy is worse when viewpoint or expression is changed between presentation and test. Bruce (1982) found that participants performed with 90% accuracy for correctly identifying an identical picture at test as one that they had viewed 15 minutes earlier. If at test the participants saw an image of the same identity that they had viewed before, but head angle or pose was changed at the test phase, recognition accuracy dropped to 76%. Changes to head angle and expression, rather than changes to just one or the other, led to further impairments in recognition performance, with performance falling to 61% accuracy. The effect of change in image between exposure and test has been shown to be so strong that even familiar face recognition slows when viewpoint is changed (Bruce, 1982).

O'Toole, Edelman & Bulthoff (1998) report a similar effect. They demonstrated that participants were less likely to recognise an image they viewed at test as an identity previously seen if the conditions of the image view (full, 3/4 or profile) had changed between the learning and test phase. When faces were learnt in full view, recognition performance was highest for test items shown in full view. However, when faces were presented in full view during the learning stage, there was an advantage for recognition from 3/4 views over profile at test. This suggests that certain changes in viewpoint might be more detrimental to recognition performance than others.

The studies mentioned above highlight that changes in viewpoint and expression influence face recognition performance. Other studies have shown that these image manipulations also affect unfamiliar face-matching accuracy (Bruce, Henderson, Greenwood et al. 1999). Bruce and colleagues (1999) found that face-matching

performance was negatively affected by any change between the target and comparison face images - this could include a change in viewpoint or change in expression. The procedure took the form of a line up scenario. Participants compared a still image, taken from video footage, to 10 face images in a line up array format shown below the target video image. The task was to determine whether the target image was present in the line up array, and if so, identify the correct match. Images in the line up array were always full face and generally neutral in expression. The still video comparison images showed either the same or different expression type (neutral or smiling) and were of either the same or different view point (full or 30 degree view) to the target image. Accuracy was worse for the trials where viewpoint or expression differed between target and comparison array images. Viewpoint change was found to reduce performance accuracy more than a change in expression.

Bruce et al. (1999) found that viewpoint could also influence performance on a paired matching task. Participants made more errors when comparing a target face to a similar distractor face if the viewpoint of the two images was different, than in situations where viewpoint of the target and distractor were the same. Hill and Bruce (1996) mention that when faces are matched across different viewpoints, performance accuracy is increased if the faces are lit from the same direction, specifically if the images are both lit from above. This is consistent with work of Johnston et al. (1992) which reported that lighting from below made it harder to recognise familiar faces. Hill and Bruce (1996) reported that changes in lighting alone reduced matching performance; as did changes in view point. Together, these findings can be taken as evidence that a change in viewpoint, and also expression or lighting, between target and comparison faces increases the difficulty of unfamiliar face-matching. Any consistencies held across the images can aid the matching effort (e.g. Hill & Bruce, 1996).

**Changes to Internal Versus Changes to External Features**

A second question of interest on the topic of which changes matter, is whether people pay different amounts of attention to *internal* and *external features* when making identity judgments. If they do, this may give us some clues to predict whether internal or external changes matter more for effective disguise. The answer appears to depend on whether the viewer is familiar or unfamiliar. Unfamiliar viewers value internal and external features as equally important when matching a face (Ellis et al., 1979; Young, et al., 1985). Familiar viewers on the other hand give more weight to internal features. Notably these are the less changeable features of a face (Ellis et al., 1979; Young et al. 1985; Tanaka & Farah, 1993; Toseeb, Keeble & Bryant, 2012). While hairstyle can change frequently, specific features of a face are generally more constant. Young et al. (1985) found these results by testing how quickly participants made correct matches for a full face image alongside only the internal features, compared to matching a full face image with an image cropped to contain only external features. Features themselves (internal and external) may be easier to compare across images taken from the same viewpoint (Bruce et al. 1999).

O'Donnell & Bruce (2001) report a slightly different but interesting finding for familiarity and attention to internal and external features. They artificially edited images of faces so that participants would see either two identical images, or the original image and an altered image of that same face side by side. Changes could be either to the hair or chin (external features) or the eyes or mouth (internal features). Participants had to work out whether the images in the pair were identical or differed physically in some way. Unfamiliar viewers (untrained) detected changes to the hair with highest accuracy (in line with findings of Bruce et al. 1999). Familiar viewers (trained) were very good at detecting changes to the hair, but were also highly attuned to detecting changes made to the eyes. This experiment made artificial changes to images resulting in an image matching type task rather than face-matching using disguise manipulations. In conclusion, familiar viewers are consistently relying on internal features of a face to aid their identity decisions. It is uncertain whether unfamiliar viewers are using internal and external

features equally (Ellis et al., 1979, Young et al. 1985), or relying more on external features (Bruce et al. 1999; 2001). Evidence can be found for both accounts in the literature.

**Past Research on What Makes an Effective Disguise**

Patterson & Baddeley (1977) begin to answer the exact question of interest, of which disguises work. As is the case for all previous disguise research, this situation was investigated for evasion disguise only. Unlike the findings above, they reported that recognition performance was not greatly affected by small changes in appearance. They specifically examined changes to pose and expression between learning and test items. Although participants showed a lower overall hit rate when the test items were of changed view and expression than identical view and expression, their false positive rate was also lower, negating any overall effect of changing viewpoint and expression. The study also tested recognition accuracy for disguised faces. Many recognition errors were made for the disguise stimuli – hit dropped rate from .98 for no change photographs to a hit rate of .45 (around chance level) for the disguise images. Disguise images in this experiment in Patterson & Baddeley's study were taken from actor photographs that showed actors in a range of different appearances depending on the roles they had taken on. Evasion or impersonation disguise itself was not directly the aim of the change in appearance. Actors had changed their appearance to suit a job role. I return to the issue of changing character later in this chapter. Active disguise manipulations, whereby someone is specifically trying to evade or impersonate identity, may be more challenging for face recognition.

Due to the type of disguise stimuli used in Patterson and Baddeley's Experiment 1, they were unable to categorize the exact changes made to appearance across images. Like me, Patterson & Baddeley wanted to understand better what appearance changes made for an effective disguise. This was addressed in Experiment 2 of the paper. In Experiment 2 stimuli were amateur dramatic students who modelled for disguise and no disguise photographs. However, unlike in my FAÇADE database, their disguises were standardised,

and were limited to the addition or removal of wigs, beards and glasses. All models were male. Participants had been informed that the appearance of the models might have been changed (through disguise) in the test images, although in Experiment 11, I saw that knowing to look for disguise did not improve performance compared to those who were not informed of the disguise manipulation. The effect of each of disguise was analysed, with the findings as follows: matching performance was greater when hairstyle remained the same than when it was changed; performance was better when a face was unchanged with regard to presence or absence of a beard; performance was poorest when multiple changes to appearance were made. The presence/absence of glasses interacted with changes in hair. The general finding was that disguise made it harder to recognise faces, and multiple disguise manipulations (more props) increased the difficulty. It is a recurrent theme that multiple alterations to a face between learning and test images have greater effect on performance accuracy than just one change, or fewer changes (Patterson & Baddeley 1977; Bruce et al. 1982; Bruce et al. 1999).

Dhamecha et al. (2014) were interested in specific elements of disguise that cause problems for face-matching performance. As explained in the introduction of Chapter 5, Dhamecha and colleagues looked at the effect of occluding facial features through the addition of props and then cropped the images so that the focus was exclusively on internal features of a face. Following collection of results from their matching task, Dhamecha and colleagues (2014) analysed performance accuracy for each segment or segments of the face covered. This was achieved by dividing the face into four regions: lips, nose, eyes and forehead. The occlusion (presence of a prop) of just one of any of these four regions was linked to high error rates in the face-matching task. However it is important to remember that due to the design type, a one-region disguise pair could consist of one face where the eyes are not visible due to sunglasses next to a face where the lips were not visible as they were covered by a medical mask (see Figure 6.1). This would leave very little of the face available for comparison across images, so it is little wonder that this is a highly error prone task. The presence of multiple props (covered areas) led to some further reduction in performance, but the greatest reduction in

performance came from the introduction of any single occlusion (i.e. from zero to one occluding disguise element).

**Figure 6.1** Figure demonstrates two different disguise images for the same identity. The image on the left occludes top part of the face, and the image of the right occludes the bottom part of the face.

All of this previous work has covered only situations of evasion disguise – manipulations that make a person look less like themselves. The situation of impersonation disguise - taking on the identity of a specific other person - has not been explored previously, thus I will be examining what makes for an effective impersonation disguise for the first time. Previous work has reported that distinguishing features are important for face recognition, with identities in a line up task being more likely to be mistaken for the true culprit if they share a distinguishing feature (Wells, et al., 1993; Flow & Ebbesen, 2007). Therefore this might provide a clue for creating convincing impersonation. Other than this, all that is predicted for impersonation disguise is that effective impersonation disguises will be those that make a face look most similar to the target face.

In summary, so far it has been established that any change between images presented at learning and test causes difficulties for recognition performance, with disguise manipulations (here seen as adding or removing props such as wigs, beards or glasses) causing more difficulty than more subtle changes such as expression and pose (Patterson & Baddeley, 1977; Bruce et al, 1992, 1999; O'Toole et al. 1998). Expression and pose

differences between images have been found to impair matching performance (Bruce, 1999), suggesting that these changes could assist the success of disguise. Disguise, in terms of occluding facial features (Dhamecha, 2014), and also free disguise (Chapter 5), leads to severe difficulties in face-matching.

In Chapter 5, I applied theories to both evasion and impersonation disguise. For an evasion disguise to be successful an individual must change their appearance so that it becomes outwith the accepted range of appearances for that face. In the case of impersonation disguise the challenge is to change appearance to get inside a specific other person's accepted range of appearances. It thus seems likely that the approach taken to create each of these types of disguise (evasion and impersonation) may be different. It is less clear whether differences will emerge between the approaches to creating the different types of impersonation disguises – impersonating someone of similar appearance compared to impersonating an identity of the same selected at random, but of the same sex. It is possible that natural similarities and differences between specific faces may guide each individual's approach, leading to a very individual method for disguise, or alternatively there may be a common approach taken by our disguise models to create impersonation disguises, for example models might consistently focus on hair or facial expression. I will approach two key questions – how people disguise themselves and which approaches are effective.

## 6.3 How do People Disguise Themselves?

### Method

### Participants

Three independent raters were instructed of the coding system and any coding discrepancies discussed until agreement was met.

**Design**

Word clouds were created for the three different disguise conditions – Evasion, Impersonation Similar and Impersonation Random.

**Procedure**

The 26 models recorded the changes that they made to their appearance by writing down on a piece of paper what they did to disguise themselves in each condition and what they hoped each manipulation would achieve. This was recorded whilst the models created their disguises.

The disguise manipulations and purposes were typed up and coded into categories to make sure the same changes, described by different words, were captured under the same word to allow a word analysis to be accurately conducted. For example, if a participant dyed their hair, this was categorised as a change of hair-colour. Categorisation was conducted by 3 raters who had been instructed of the coding system. Any discrepancies were discussed amongst the raters until agreement was met.

Hyphenations were used to make sure that meaningful phrases, such as hair colour, were kept as one word. Otherwise the word *hair* for example, would have been counted across several different hair changes - such as hair colour and hair length. Hyphenating words that contributed to a phrase ensured accurate word categorisation.

The categorised body of text was entered into wordle (http://www.wordle.net) – software that creates *word clouds* based on word frequency. Words are given prominence within their cloud based on the frequency of their occurrence. The more often a word occurs in the inputted text, the larger the word will be displayed in the word

cloud. These word clouds may help in gaining knowledge of the manipulations made by the models for each of the disguise conditions.

**Results and Discussion**

**Evasion**


**Figure 6.2** Word cloud showing the most frequently stated words for creating an Evasion disguise.

All changes in the Evasion condition related to creating differences in appearance compared to the model's own reference photograph. Models frequently changed the appearance of their own features, skin-tone, hairstyle or hair colour and clothes in order to look make themselves look physically different to their normal (own reference photograph) appearance (see Figure 6.2). The word cloud includes ways that participants tried to change their identity and shows that this was largely done through props, such as use of the words 'wig', 'glasses' and 'clothes'. Props were however not the only method used to create facial change. Other techniques used included the use of makeup to facilitate changes to features and skin tone and changes in camera angle to change face shape.

**Impersonation Similar**


**Figure 6.3** Word cloud showing the most frequently stated words for creating an Impersonation Similar disguise. All words represent similarities with the target face except where specified as differences.

To create Impersonation Similar disguises models focused on looking similar to, or even matching the appearance of, the reference photograph. This was attempted in a variety of ways including focusing on creating similarities in the appearance of specific features especially hairstyle but also eyes and lips (see Figure 6.3). Beards were grown or shaved where applicable to match the condition of the reference photo (see Figure 6.4). Copying the eyebrows of a face was a particularly common impersonation similar disguise technique (see Figure 6.5). This was done by changing the eyebrow colour, covering up eyebrows with foundation, or drawing on eyebrows in a different style or shape to the model's own to create a better eyebrow match with the reference photo. Copying the facial expression of that in the reference-photo was also a very popular impersonation technique (as in both Figure 6.4 & 6.5).


**Figure 6.4** The model (shown right in this image pair) shaved his beard to better match the appearance of his target (left).


**Figure 6.5** The model (right) has copied the eyebrows of the target (left) using makeup to alter eyebrow shape.

**Impersonation Random**


**Figure 6.6** Word cloud showing the most frequently stated words for creating an Impersonation Random disguise. All words represent similarities with the target face except where specified as differences.

For the Impersonation Random condition, models took a roughly similar approach to their disguise as they did in the Impersonation Similar condition. Models again focused on creating similarities with the reference-photo and matching on a range of features. It is interesting to note that hair, an external feature, became the predominant focus of change, and there was slightly less focus on internal features (see Figure 6.6). The changes to internal and external features were somewhat more balanced in the case of impersonation similar disguises. Perhaps models felt it harder to manipulate internal features when there were less natural similarities, additionally there could be more differences between hair style in the random matching to account for. There was also a little less focus on copying facial expression than there was in the impersonation similar condition. This may be due to expression matching resulting in a picture matching type effect, whereby the impersonation random models could be revealed as an impersonator if viewers spot slight differences in the copied pose.

The most common approach to impersonation disguise fits with the theoretical framework adopted in this thesis. Impersonation disguises rely on successfully creating a face image that would be accepted as falling within the other person's face space. The

models make a clear effort to do this, as they focus on creating similarities with the reference photograph. It is unknown at this stage whether the changes that were most common were actually those that caused most difficulties for matching. This will be looked at in section 6.5. First I will discuss whether viewers can tell by eye which disguises will be effective.

## 6.4 Can Viewers Predict by Eye which Disguises will be Effective?

The next question of interest is whether the images rated to be the best disguises were those that caused most difficulties in the matching task. This investigation will help to understand if viewers can determine disguise effectiveness simply through a side by side comparison of the target and disguise face images (evasion, impersonation similar, impersonation random).

### Method

#### Participants

The independent viewer group (first mentioned in Chapter 5), who chose the final stimuli for the FAÇADE database, also provided the participant group for this study.

#### Procedure

The independent viewer group provided disguise effectiveness ratings for each of the selected database disguise images on a scale from 1-7 with 1 being a very poor disguise and 7 an extremely convincing disguise.

**Design**

To find out whether the viewer group could accurately predict disguise effectiveness, I compared the effectiveness ratings for each of the disguise images as scored by the group of four unfamiliar viewers, to performance accuracy for each of the items on the face-matching task. This was conducted for each of the three disguise conditions – Evasion, Impersonation Similar and Impersonation Random.

**Results**

Pearson correlation coefficients between ratings of by disguise effectiveness of the face images and performance accuracy for the disguised images on the face matching were calculated to find out whether effectiveness ratings predicted matching accuracy.

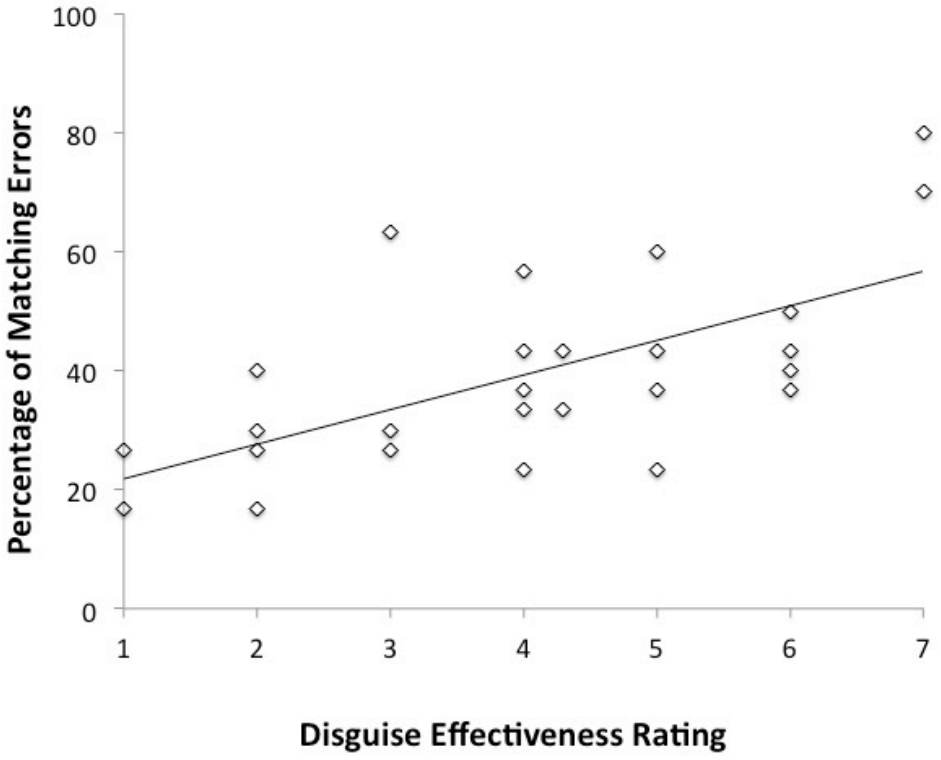


**Figure 6.7** Graph showing correlation between effectiveness rating and percentage of errors made for each Evasion disguise item. Data points are spread horizontally if they would otherwise overlap.

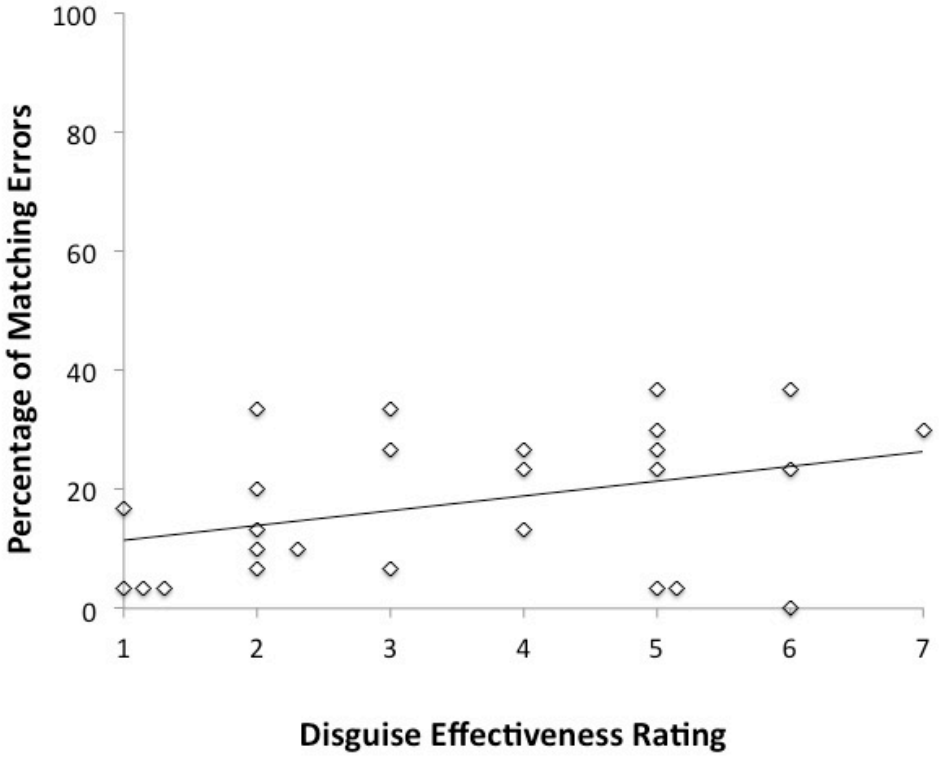


**Figure 6.8** Graph showing correlation between effectiveness rating and percentage of errors made for each Impersonation Similar disguise Item. Data points are spread horizontally if they would otherwise overlap.

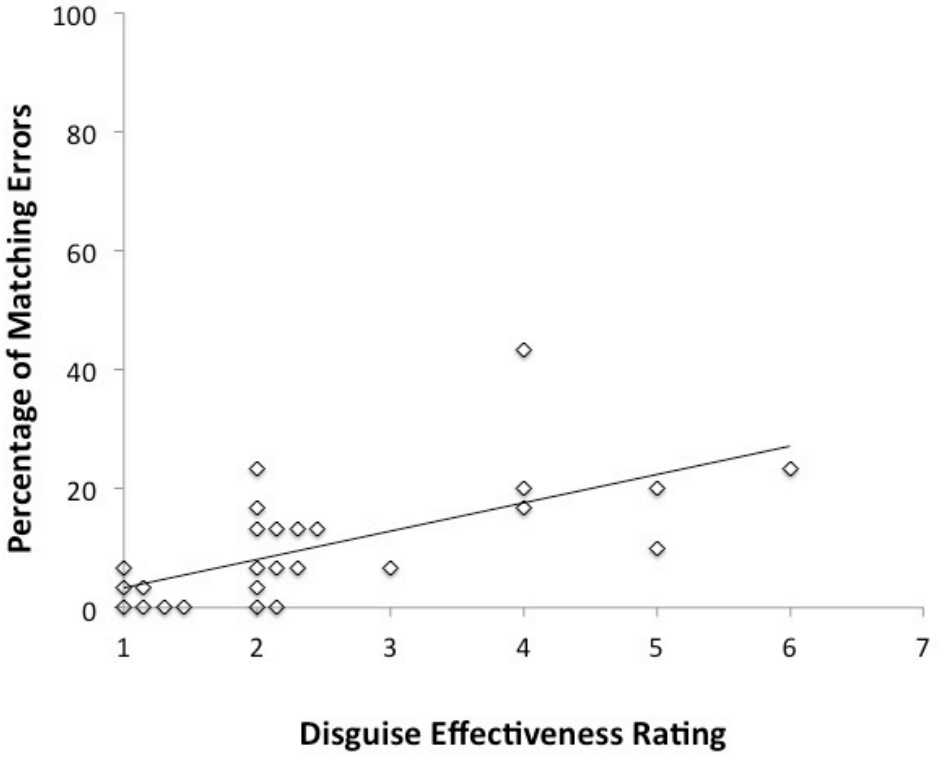


**Figure 6.9** Graph showing correlation between effectiveness rating and percentage of errors made for each Impersonation Random disguise item. Data points are spread horizontally if they would otherwise overlap.

Disguise ratings were strongly correlated with performance on the face-matching task for both Evasion disguises r =.66, p<.01 (see Figure 6.7) and Impersonation Random disguise r = .65, p<.01 (see Figure 6.9). For these conditions the higher the effectiveness rating of the disguise, the more face-matching errors were made. Correlation levels were almost significant in the disguise impersonation similar condition, r = .39, p=.05 (see Figure 6.8). The Impersonation Similar condition is naturally a more difficult condition for the raters to judge based on the similarity of the target and reference photograph. There was however a highly significant correlation for impersonation when collapsing across impersonation type, r = .54, p<.001.

These results suggest that the effectiveness of a disguise can be judged quite accurately by simply showing a group of viewers both the reference photograph and disguised image side by side and asking them to rate how good the disguise is.

## 6.5 What do Viewers Believe Makes for an Effective Disguise?

Now I have shown that viewers can accurately predict which disguises would be effective, I believe it is interesting to explore what it was about a disguise which the viewers thought made it effective. It is believed that as the viewers were accurate at making their disguise effectiveness decisions, their insight into what made an effective disguise will also be useful. Obtaining an understanding of what makes an effective disguise may make it easier to create successful disguises in the future; this could have useful applied value such as aiding the disguise of undercover police officers.

## Method

### Participants

The participants were again the 4 raters who made the stimuli selection decisions discussed in Chapter 5, and also gave the disguise effectiveness estimates used in the study above.

### Design

Effective disguise changes were categorised and counted for each of the three disguise conditions.

### Procedure

As part of the stimuli selection phase, four unfamiliar viewers worked collaboratively to decide upon the best match or mismatch image from a range of options for each model for each disguise condition. During this process I asked the viewers to rate how good they thought that each of the disguises were for each of the selected images on a scale from 1-7. A score of 1 indicated that the disguise was very poor, whereas a score of 7 meant that the disguise was extremely effective. These were the ratings used for the correlations in section the study above. In addition to these ratings, the viewers provided comments on what they thought made the chosen disguise images effective in each of the conditions.

The comments made by the viewers were coded into seven categories of change – internal features, external features, expression, skin-tone, social inferences, face shape and other.

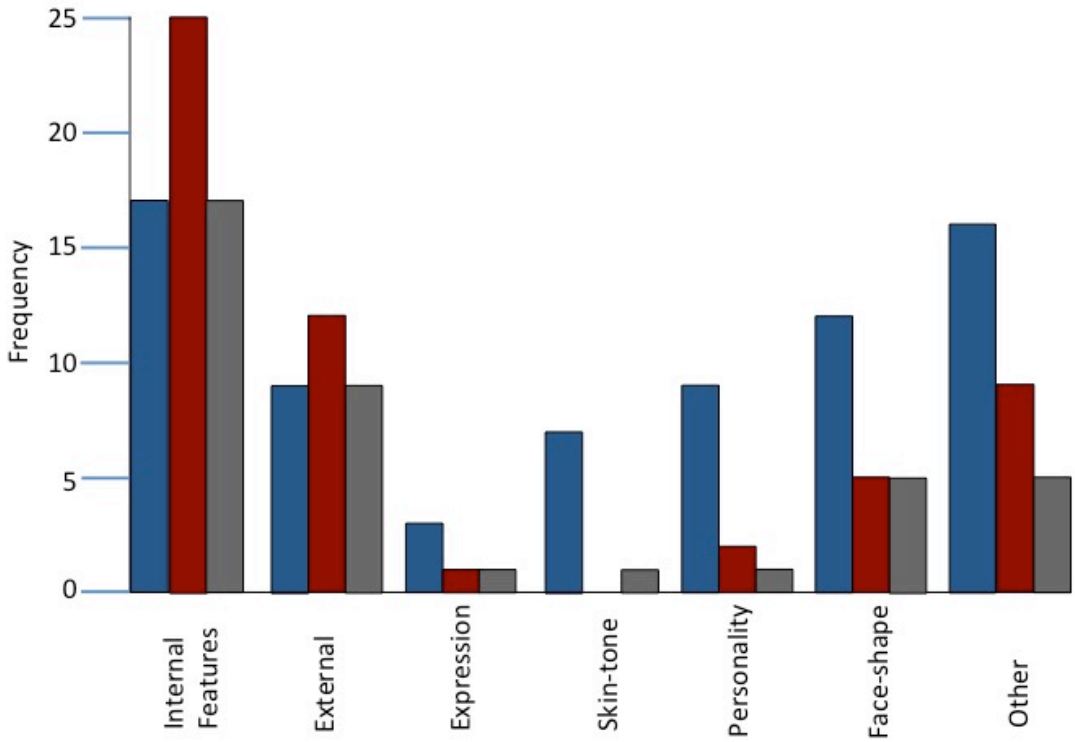


**Figure 6.10** Bar graph showing the most frequent forms of disguise for Evasion (blue), Impersonation Similar (red) and Impersonation Random (grey). Evasion changes capture differences in appearance with the reference photograph whereas Impersonation changes represent similarities.

Figure 6.10 shows the frequency of effective changes by category for each of the three disguise types. From this figure it is evident that effective Evasion disguise was created by many different types of changes. Changes to Internal (specific features of the face) and External changes (changes to the hair and also any addition of props, including clothing) were made to the face (specifically changes to nose shape, eyes and hair), but changes in Expression, Skin-tone, Personality, Face-shape and Other changes (e.g. changes in lighting) also contributed greatly to effective Evasion disguise. Whereas the models tended to focus mostly on changing hair and makeup when creating their disguise (Figure 6.2), interestingly the viewer group also picked up on the personality differences and changes to face-shape that these changes evoked and found these to be effective factors

of change to create evasion disguise. For example, viewers described a disguise as effective if they thought that the images represented different characters, such as one image showing a quiet and studious person and another showing a more outgoing and party loving person (see Figure 6.11).

**Figure 6.11** Image example where social inferences were reported to differ between the reference model image (left) and the model in evasion disguise (right).

Using the numerical information entered into the bar chart above, a binomial test was conducted to find out whether there was a significant difference in the proportion of Feature to Non-Feature changes for Evasion disguise compared to Impersonation disguise. The test proportion was calculated from the proportion of Feature (represented as internal features and external on the graph) and Non-Feature (all other categories of change from the graph) changes made for Evasion disguise, and this was compared with the same proportions for Impersonation disguises. The binomial test indicated that for Impersonation Similar disguises the proportion of .69 feature responses was higher than the expected .34, $p < .001$. This same pattern of results was found for Impersonation Random disguises, with the proportion of .67 features being higher than the expected .34, $p < .001$.

These binomial results highlight that whereas effective Evasion disguise encompassed many different changes, across many different categories, with other changes

outweighing internal and external changes, effective Impersonation Similar and Impersonation Random disguises were dominated by the internal and external feature changes. The bar graph shows that internal featural similarities were the most common effective change for impersonation similar disguises (similarities between the nose, ears, eyes, forehead, eyebrows & mouth). Changes to external features followed, these were generally related to hair. Personality and expression were much less noted for creating an effective impersonation than they were for evasion. These results suggest that to create an effective impersonation similar disguise the focus should be on creating similarities between the internal features of a face. The effective Impersonation Similar manipulations were very similar to the effective Impersonation Random manipulations. Internal features again came out as most important, followed by hairstyle changes (External). For the impersonation conditions, models sometimes tried to copy any distinguishing feature of the person they were trying to impersonate. This made the disguise more effective and was captured under the category 'Other' (see Figure 6.12 for example). The presence of distinguishing features has been found to affect face matching accuracy in past studies of facial recognition (Wells, et al.,1998).

**Figure 6.12** The model (left) copied the distinguishing feature (mole [on the left side of the image under the mouth]) of the target (right) by using make up.

More changes, and categories of change contributed towards effective Evasion disguise than Impersonation disguise. Effective Impersonation disguises generally revolved around changes to internal and external features of a face whereas Evasion disguise also included changes in perceived face-shape and personality, as well as 'other' manipulations, which were those not captured by the specifically labelled categories. It is important to remember that previous research has been limited to looking at effective Evasion disguise only, and even within that Evasion disguise was investigated in relation to the effect of specific props and occlusions of certain areas of the face. No features were occluded in my disguise database. Previous studies have missed most of the action by focusing only on the category that I named External. My results show that far more goes on in the creation of free disguise, for both Evasion and Impersonation, than simply the addition of props, such as wigs and glasses.

## 6.6 Experiment 13 - Do Social Inferences Change for Disguise?

The graph and binomial analysis above demonstrated that the viewers' disguise effectiveness verdicts were influenced not only by the featural types of change explored in previous research, but also by aspects relating to social inferences. This was especially true in the case of evasion disguise. I believe this is an interesting topic to investigate further – most research on disguise has used props to change appearance. Never before in the literature has disguise been studied with the intention of investigating disguise related changes in perceived personality traits.

Patterson & Baddeley (1977) showed that encoding faces in terms of social inferences was a powerful tool to aid facial recognition. In their experiment participants viewed face images during a learning phase and were instructed to make judgments on either the features or personality of the person in the image. Participants who were assigned to the feature condition made ratings of nose size (small – large), lip size (thin – full), width between eyes (close together – far apart), face shape (round – long). Those instructed to make personality judgments rated niceness (nice – nasty), reliability (reliable –

unreliable), intelligence (intelligent – dull) and liveliness (lively – stolid) of the individuals shown in the images. In both of the conditions participants were informed that making these judgments of the face might help them to remember the faces. It was found that those participants who made personality trait inferences for the faces were better (but not significantly so) at recognising faces at test than those who made feature related judgments. This trend provides an interesting hint that personality perception may influence the encoding of faces, but cannot be taken as evidence as a significant result was not found. The sample size was fairly small at 18 participants in each condition; a larger sample size may have provided a significant result. Combined with the comments received from my viewer group regarding what makes a good disguise, it is possible that changes in social inferences may have influenced performance on the face-matching task of Chapter 5.

Trait perception research has generally considered whether *accurate* social inferences about a person can or cannot be acquired from viewing a photograph of that person's face. Those studies have typically considered natural images, meaning that the photographed person has not been asked to express any particular trait, but instead may or may not reflect aspects of their true personality unintentionally in photograph. To find out whether viewers' judgments of personality are accurate, scores from personality questionnaires taken by the models are generally compared with viewers' trait scorings of the photograph. Several studies have found that above chance level judgments can be made in this way. For example extraversion was accurately judged from face images viewed for as little time as 50ms (Borkenau, Brecke, Möttig & Paelecke, 2009). This social judgment was linked to the display of joyful facial expressions such as smiling (Kenny, Horner, Kashy & Chu, 1992; Borkeneau et al. 2009). Personality judgments recorded by Naumann, Vazire, Rentfrow & Gosling (2009) were accurate in both an uninstructed natural photograph and a neutral pose photograph condition, with the authors reporting that expression was not the only factor to aid social inferences from a photograph. Factors such as clothing were reported to also lead decisions, as well as general appearance and the way people held themselves. Rule & Ambady, 2011 even suggest that earlier life photographs can suggest success in later life, with power inferences from

college yearbooks predicting later leadership success. These findings portray social inferences as a stable concept attached to an individual, which holds across photograph and time. If this is the case, then social inferences could actually help in identity judgment tasks.

Whereas most past research looked at the ability to take accurate social inferences from photographs and perceive these inferences as a stable judge of character, in some situations people may want others to infer inaccurate social inferences from their photographs. For example appearing to be a better candidate for a job or enhancing likeability on a social media account or online dating profile. Leikas, Verkas & Lonnqvist (2012) found that it was possible for the same person to have different social inferences made about them across different photographs. Models were asked to change in appearance to match high and low ends of the spectrum on each of the Big 5 Personality Traits, and photographed in each of these conditions – therefore the study focused on deliberate changes in appearance rather then incidental change between different photographs of the same face. Each of the Big 5 Personality traits was described by two adjectives that the models could use as a guide for creating their appearance in the corresponding photograph. For example 'anxious and stressed' described the trait neurotic, whereas in the stable condition models were asked to make themselves look 'stable and quiet'. Models were limited in the ways that they could achieve these personas, with the paper stating that 'targets were not allowed to add, change or remove clothing, hairbands or decorative items, to remove or add makeup, or to groom their hair between conditions' (Leikas et al. 2012). Extroversion could be convincingly changed across photographs, and to a lesser degree neuroticism and conscientious. Agreeability was not successfully changed across posed photographs. These findings suggest that it is possible to deliberately control social inferences in a photograph. I may see even greater changes in the disguise experiment as my disguise models were given far greater freedom regarding what they could do to change appearance, and as shown in the section above, the models used a great range of strategies to change their appearance in addition to purely pose and expression.

It is possible, and seems viable based on the findings of section 6.5, that social inference may change as a result of disguise manipulations. Some incidental change is expected across multiple undisguised images of the same face. In this next experiment, I will test whether social inference judgments differ more for disguise change than naturally as a result of incidental change. The traits that will be explored are Trustworthiness, Attractiveness and Dominance. These traits have been chosen as they are the 3 traits that people spontaneously characterise faces on, according to PCA analysis (Todorov, Said, Engell & Oosterhof, 2008; Sutherland et al. 2013).

For the case of evasion it is predicted that there will be a greater difference in social inferences for disguise related change (different in social inference judgments between the reference photograph and disguise photo) than for incidental change (difference in social inferences between two undisguised images of the model). As models creating evasion disguise are trying not to look like themselves in their evasion disguise image, creating a situation where their social inferences are changed in relation to those made for a no disguised reference image may help to hide their true identity. For impersonation disguise models tried to look more similar to the target, this may have been reflected in similar social inferences. For the case of impersonation disguise it is predicted that differences in social traits will be larger for incidental change (the target photograph and undisguised model image) than between the impersonation disguise image and the image of the target person).

**Methods**

**Participants**

30 undergraduate students (M = 12, mean age = 20.4) from the University of York (who had not taken part in any previous experiments involving the FAÇADE data set)

volunteered as participants in this experiment in exchange for a small cash reward or course credit.

**Stimuli & Design**

The stimuli were all the images from the FAÇADE image database, presented one at a time. For each model the following images were shown: reference, no disguise, evasion disguise, impersonation similar disguise, target similar person's reference, impersonation random disguise and the target random person's disguise reference image. All images were presented in a random order.

This is a within subjects design whereby each participant rates each face image on each of the three Traits (Trustworthiness, Attractiveness and Dominance). Each of the Traits were rated in a random order for each face image.

**Procedure**

Participants viewed the stimuli (one image at a time) on a computer screen. Underneath each image appeared the question, either 'How trustworthy is this person', 'How dominant is this person?' or 'How attractive is this person?' The image remained on screen until all 3 of these questions had been answered for the face, just one question was shown at a time and the questions were always asked in a random order. Participants indicated their response to each question by using the computer mouse to select an answer on a rating scale between 1 (low score on the trait) and 7 (high score on the trait).

**Results and Discussion**

Social inference distance change was always measured according to the square root distance between face images rather than the distance between means. Square root distance took into account the fact that some people naturally looked more similar than others did to the person they were trying to impersonate on certain traits. The square

root distance was calculated for disguise to target face images, and compared to the incidental change for the corresponding no disguise images.

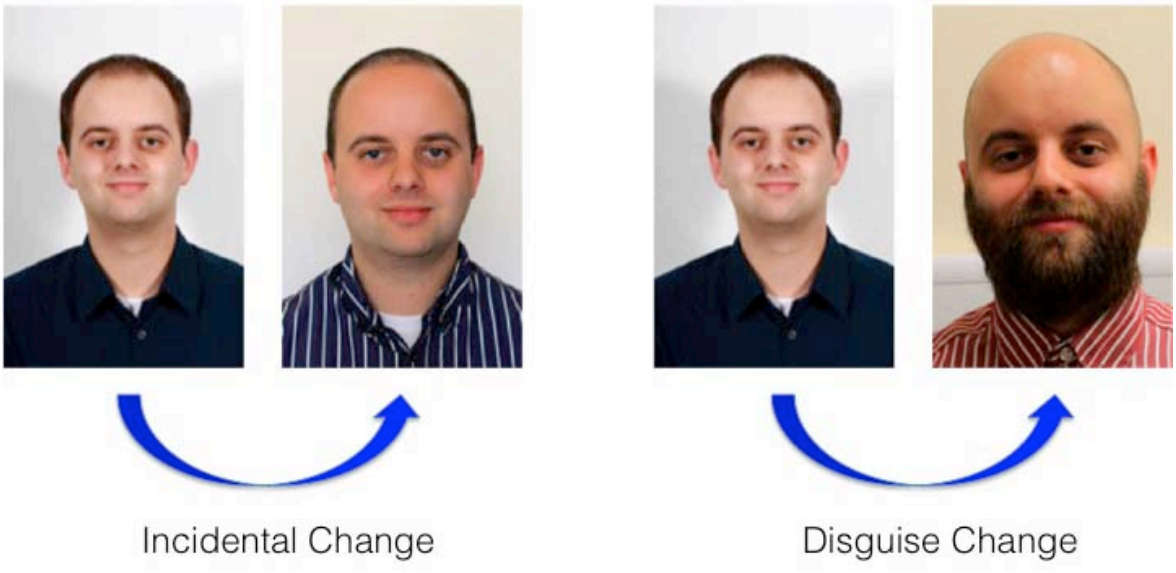


**Figure 6.13** Example illustration of the distance calculations made for the Evasion disguise condition. Distance moved for incidental change was compared with distance moved for disguise change for each of the 3 disguise conditions (Evasion, Impersonation Similar, Impersonation Random).

In cases of evasion people moved significantly further from their perceived personality traits (reference image) compared to the change in trait perception for another no disguise image of that person, $t(25) = 7.71$, $p < .001$, CI = 1.19 to 2.06. The mean distance moved in trait perception between reference photo and no disguise image was 0.63, SD = 0.38, SE = 0.08. The mean distance moved in trait perception between reference photo and evasion photo was 2.25, SD = 0.92, SE = 0.18.

To investigate where these differences in traits lay, results were analysed as above, but this time for each social inference (Trustworthiness, Dominance and Attractiveness) in turn. There was a significant difference in perceived trustworthiness scores between the reference and no disguise self images M = .34, compared with the difference between the reference and evasion images M = 1.43, $t(25) = 6.54$, SE =.16, CI = 0.74-1.43, $p<.001$.

There was also differences in perceived dominance scores between reference and no disguise self images M = .24, compared with the difference in perceived personality traits for the reference and evasion images M = .60, t(25) = 3.55, SE =.10, CI = 0.14-0.56. p<.005. Finally, there was a significant difference in perceived attractiveness scores between the reference and no disguise self images M = 0.36, compared with the difference in perceived personality traits for the reference and evasion images M = 1.42, t(25) = 5.63, SE =.19, CI = 0.67-1.44, p<.001. These results show that the there were significant differences in social inferences for each inference in turn as well as an overall effect. This was as expected based on the results of the effective disguise section earlier, where personality change was listed as an effective form of disguise.

**Impersonation Similar**

For impersonation similar faces there was not a significant difference between perceived social inferences as rated as a whole for the reference and impersonation images (Mean distance = 1.19), compared with the difference in perceived personality traits for the reference and no disguise self images (Mean distance = 1.24), t(25) = .38, CI = -0.31 – 0.21, p =>.05, SE = .13. This was the case for each inference individually and overall (see Table 6.1).

| | Distance between target reference image & impersonation image | Distance between target reference image & model's reference image | Significance Level |
|---|---|---|---|
| All traits combined | 1.19 | 1.24 | p>.05 |
| Attractiveness | .73 | .82 | p>.05 |
| Dominance | .38 | .42 | p>.05 |
| Trustworthiness | .67 | .60 | p>.05 |

**Table 6.1** Social inference comparisons for impersonation similar images.

**Impersonation Random**

The pattern of results for impersonation random items was similar to that for impersonation similar items – no results were significantly different at either an overall level or trait breakdown (see Table 6.2).

| | Distance between target reference image & impersonation image | Distance between target reference image & model's reference image | Significance Level |
|---|---|---|---|
| All traits combined | 1.31 | 1.24 | p>.05 |
| Attractiveness | .75 | .97 | p>.05 |
| Dominance | .51 | .55 | p>.05 |
| Trustworthiness | .75 | .66 | p>.05 |

**Table 6.2**. Impersonation random disguises, means and median results for each analysis.

The results for impersonation disguises were not as hypothesised – impersonation disguises did not move a target face significantly closer to the trait perceptions of the model. This is perhaps not entirely surprising, as similar personality was not picked out as particularly important in the effective disguise section above. It seems that in the case of impersonation, other factors, such as featural change may be more important than changes of social inference.

**Discussion**

Social inferences differed significantly in the case of evasion disguise, suggesting that a change in these perceptions occurs when people are trying to not look like themselves. Social inference must be based on physical appearance. The physical appearance changes

made to create Evasion disguise influenced social inferences. Judge of character is generally believed to be accurate, but people are fooled by social inference change in terms of identity judgment in the case of evasion disguise.

Impersonation disguise on the other hand does not rely on the match of social inferences between a model in impersonation disguise and an image of the target person that the model was trying to impersonate. In this scenario other factors of similarity seem to be more important e.g. specific and distinguishing feature match. Instead of trying to mimic the target *image* itself, e.g. the exact pose and expression, models may instead be attempting to portray another way that the target identity could appear, e.g. a different pose and expression. It has already been established that it is harder to match identities over changes in expression (Bruce et al. 1982), than in unchanged expression. This may work to an impersonator's advantage in the case of disguise, and also explain why social inferences are not matched for impersonation similar disguise. If for example, a model chooses to try and look like the target person when they are in a happier or angrier mood, the social inferences drawn from the face may differ even if changing the expression makes the images look more similar in identity overall.

## 6.7 General Discussion

In summary, the disguise models applied many different manipulations to create each of their disguises. Viewers were able to predict, by looking at a target image and disguised comparison image, whether the disguise was effective or not (with relation to the percentage of errors made for that face in the face-matching task Experiment 10). Furthermore, manipulations that make an effective evasion disguise are different to those that make an effective impersonation disguise. Whilst all disguises involve either the similarity or difference in internal and external features, evasion disguise also encompassed more non-feature factors including changes in expression and perceived personality. Finally, Experiment 13 confirmed that social inference change occurred only in the case of evasion disguise.

Evasion disguises involved internal features and external changes, with internal changes often being achieved through the use of makeup and also changes in expression. Impersonation similar changes involved creating similarities with the target face which included similarities in hairstyle, internal features, clothes and expression.

Impersonation random disguises focussed mostly on creating similarities in external features. The most important message here, with relation to previous work, is that models did far more to create their disguises than simply the addition of props. Prop additions and occlusion of features with props have been the most common method of creating disguise stimuli for disguise investigation in previous research. My research shows that when given the freedom to create their own disguises, models use many more disguise techniques other than simply the addition of props. If disguise research is to have real world relevance, then free disguise manipulations should be allowed when constructing model stimuli.

Additionally, this chapter reiterates the differences between Evasion and Impersonation disguise. Chapter 5 highlighted that Evasion caused more matching difficulties than Impersonation. This chapter goes further by showing that the approach taken to these disguises, and the factors which make them effective, differ for evasion and impersonation. Whereas both disguise types include featural manipulations made by the models, the independent viewer group picked out more changes in personality and expression as factors that made a disguise effective for Evasion disguise than in either case of Impersonation. In line with this finding, social inference change occurred only for Evasion disguise. These results further highlight the importance of a distinction between evasion and impersonation when investigating disguise.

One interpretation of these findings could be that large differences in appearance tend to make ‘same’ responses unlikely and be accompanied by changes in social inferences. Because Evasion disguises are less constrained than Impersonation disguises, they can result in major image differences. That leads to higher error rates (as seen in Chapter 5) and distinct social attributions.

It seems that effective disguise does not revolve around a simple recipe that would work for all faces. Lots of different changes are taking place, and some of these changes will work better for some individuals than others depending on the natural appearance of a face. What is evident is that there is more to disguise than simply the addition of props and furthermore Evasion and Impersonation disguise are achieved in different ways.

# Chapter 7 – General Discussion

## 7.1 Overview of Findings

The research reported in this thesis investigated face recognition in challenging situations. The introductory chapter outlined that the critical task of face-matching is difficult for unfamiliar viewers, yet trivial for familiar viewers. These findings came from past studies that used cooperative stimuli, meaning that for *same* face pairs, the person who was photographed made no deliberate attempt to change their own appearance across multiple images, and images were taken under constant conditions with extremely short time intervals between photographs. In cooperative stimuli tasks, *different* face trials paired the most similar faces from a small pool of available images (Burton et al, 1999; Bruce et al. 1999). In this thesis I argued that performance is likely even worse for images of a challenging nature – images that include incidental or deliberate face variation or across identity similarities and reduced image quality. I explain that there may even be limits to the familiarity advantage (times when familiarity can not completely compensate for poor performance). I investigated face-matching situations where the same face image pairs incidentally looked different due to within person variation across ambient images, and also where different identity pairs were of extremely similar appearance due to natural facial similarities between celebrity and lookalike faces (Chapter 2). Investigation continued for unintentional appearance change due to change of camera-to-subject distance (Chapter 4). And finally I examined matching performance for deliberate disguise, where an individual deliberately made changes to evade their own identity or to impersonate someone else (Chapter 5). The manipulations made to create the disguise face-matching stimuli were explored (Chapter 6). I also tested ways of improving performance for challenging face images (Chapter 3). In total, 16 studies were conducted with the aim of furthering understanding of face recognition in challenging situations.

The investigation began by testing face-matching performance for images in which different identity pairs consisted of faces that were extremely similar to each other and

same identity pairs included naturally captured variation within a face (Chapter 2). Celebrity lookalike images were used as naturally occurring imposter faces, and these images allowed the creation of difficult different face pairs (one image of the celebrity and one image of a lookalike for that celebrity) which appeared in a face-matching task along with same identity pairings (two different [ambient] images of the same celebrity's face). Participants were instructed to make same or different identity judgments for each of the image pairs. Experiment 1 demonstrated a graded effect of familiarity for the lookalike task - unfamiliar viewers made many identity matching errors, whereas familiar viewers performed with near perfect accuracy. My pattern of findings in terms of familiarity and performance demonstrated lower accuracy on my celebrity lookalike task (mean performance accuracy = 72%) than performance for unfamiliar viewers on the GFMT, a standardised face-matching task, which contained cooperative stimuli (mean performance accuracy long version = 89.9%, short version [hardest 40 items] = 81.2%). To model the applied problem of varied image quality, Experiments 2 and 3 proceeded to make the task harder still by reducing image quality through pixelation. The images were degraded making them challenging but also realistic of the image type often acquired from zoomed in digital images. A graded familiarity advantage survived through Experiment 2, however for the highest pixelation level (Experiment 3) performance was around chance for all but the extremely familiar viewers. These findings highlight firstly that performance for challenging stimuli image matching is even worse than the poor performance already established for cooperative face image matching. Additionally, the findings highlight familiarity as a graded concept. Finally my findings suggest that the familiarity advantage has limits – only extreme familiarity with a face could help in the case of coarse images, and this advantage remained only for same identity trials (Experiment 3).

I found that image quality affected overall face-matching performance, and the influence that familiarity could have on performance. Matching accuracy was lower for degraded (mid-pixelated and coarsely pixelated) versions of the celebrity and lookalike images (Experiments 2 & 3) than for the un-manipulated (fine quality) versions of the images (Experiment 1). In forensic investigations, the images that are available for comparison

often have this pixelated appearance. It was therefore important to investigate ways of improving identification accuracy for these poor quality images (Chapter 3). Chapter 3 explored whether techniques that have been successful in improving cooperative face-matching accuracy, could also improve accuracy for challenging images. Blurring the pixelated images (Experiment 4), pooling judgments via crowd analysis (Experiment 5) and superior ability of super-recognisers (Experiment 6) all improved face-matching performance for challenging images. I also found that some of these techniques could be combined for additional benefits.

I next looked at a naturally occurring image manipulation – change of camera-to-subject distance – to investigate how this change affected facial appearance for individuals across photographs, and whether such changes in appearance caused difficulties for face-matching. Thus, this was an investigation of an unintentional change to appearance, which may in turn influence identity judgment. Measuring face images (Experiment 7) showed that changing camera-to-subject distance resulted in non-linear changes in distances between features of a face, such that ratio measurements between features were not preserved when camera-to-subject distance was altered between photographs. Experiment 8 showed that these differences in facial configurations across images caused matching difficulty for unfamiliar viewers. Performance for same identity pairs was poorer when matching across images of varied camera-to-subject differences, compared to performance for images taken from the same camera-to-subject distance. Familiar viewers were unaffected by the camera-to-subject manipulation. I also showed that viewers compensate for changes to camera-to-subject distance when distance cues are available (Experiment 9), implying a high-level perceptual constancy for face shape.

The focus of investigation moved next to intentional appearance changes – deliberate disguise. I created the FAÇADE image database as a resource that includes evasion and impersonation disguise stimuli as well as undisguised comparison images of the same identities. These images made it possible to perform a direct comparison between disguised and undisguised face-matching performance. I found that deliberate disguise

impaired face-matching for unfamiliar viewers (Experiment 10), even when participants were informed of the disguise manipulations (Experiment 11). Not all disguises caused equal difficulty. Evasion impaired performance more than impersonation. Moreover matching accuracy was higher when the target and model's impersonations were based on random matching than when they were of naturally similar appearance prior to the impersonation disguise. Interestingly, familiar viewers performed better than unfamiliar viewers overall (Experiment 12), but they too were worse at matching evasion faces than undisguised face pairs. Chapter 6 was more exploratory, and investigated how the models in the FAÇADE database disguised themselves and what made for an effective disguise. A social inference experiment concluded the investigation to find out whether the disguise manipulations applied affected the social inferences made for the individual. Viewers made significantly different social trait inferences about the models, when they viewed an undisguised image of the model's face than when they viewed the same model's face in evasion disguise. However, for the impersonation scenario, social inference ratings were not significantly more similar between images of the target face and impersonation face, than between the undisguised model image and impersonation image. This suggests that whereas social inference related changes are important for creating evasion disguise, other factors are more important for creating impersonation disguise.

In summary, the results of this thesis show that face-matching performance can be impaired both by incidental changes in appearance (Chapters 2 & 4) and by deliberate changes in appearance (Chapter 5). There are several methods of improving this poor performance (Chapter 3). In some instances natural solutions can be exploited, including taking performance from those with high face recognition aptitude (super-recognisers, Experiment 6) and high-level perceptual constancy can account for camera-to-subject distance related face changes (Chapter 4). Familiarity improved performance throughout. I explored what people did to disguise themselves, and what works. Disguise is more complex than previous studies have allowed for. For example evasion and impersonation result in different levels of difficulty and the disguises themselves were created using different techniques and facial manipulations. Effective disguise involves more than the simply the addition of props, but it is not the case that there is a simple disguise formula

which can be applied to all – some disguise manipulations and combinations work for some individuals but not for others. Not all questions relating to disguise could be answered within this work.

## 7.2 Relation to Previous Research

Previous research had already established that unfamiliar face-matching performance is poor. All of my experiments support this, as I consistently found poor performance for unfamiliar viewers (Chapter 2, Chapter 4, Chapter 5). My findings add to this past research by showing that accuracy for challenging (but realistic) image conditions is lower than for cooperative face-matching, something that was already known to be poor (e.g. Burton et al. 1999). In support of previous findings, my results show much better performance for familiar viewers than unfamiliar viewers in each of my face-matching tasks (Chapter 2, Chapter 4, Chapter 5). I also found a graded familiarity effect for face-matching performance similar to that demonstrated by Clutterbuck & Johnston (2001, 2003), and implemented a new familiarity scale to capture this concept (Experiment 1). There are also some important differences between my findings and past familiarity advantages. In past research familiarity has generally led to performance to be at ceiling (e.g. Burton et al. 1999; Bruce et al. 2001), and this has made it difficult to quantify changes in performance. The challenging nature of my tasks took familiar viewer performance off ceiling level (Chapter 2, Chapter 5), revealing important gradations at the end of the range. Furthermore, limits of familiarity began to become clear in Experiment 3, when familiarity was tested to destruction. The graded nature of the familiarity advantage ceased to exist for the coarse image version of the lookalike task. However the advantage did not break down completely; extremely familiar viewers outperformed the unfamiliar viewers at the task, but this was only true for same person trials and may reflect the greater variation of appearances held for familiar faces, accessed when the identity of the target celebrity is recognised by the viewer. Familiar viewers performed the same task with high levels of accuracy for both same and different face image pairs when the image quality was good (Experiment 1) or slightly degraded (Experiment 2) and past studies report that familiar viewers remain unaffected by image degradation (Burton

et al. 1999). These results thus suggest that the combination of the similarity of faces and severely degraded image quality made the task difficult for familiar viewers, rather than either of these factors taken in isolation.

Although I began to find limits of familiarity in Experiment 3, it was in Experiment 12 that I saw the performance of familiar viewers suffer due solely to the model images being challenging, rather than due to challenging images combined with degraded image quality. In Experiment 12, I found that for evasion disguise faces, familiarity improved performance compared to performance for unfamiliar viewers, but familiarity did not completely compensate for the effect of disguise. Familiar viewers made more errors when matching evasion disguise face pairs than undisguised versions of these pairs. The results of Experiments 10, 11 & 12 demonstrate that the evasion of identity created a more difficult face-matching scenario than impersonation of identity. The scenario in Experiment 12 where familiar viewers performed with lower accuracy for disguised and undisguised faces occurred for evasion disguise only. Indeed evidence from other experiments in this thesis is consistent with the deliberate disguise findings. The celebrity lookalike images presented in Experiment 1, for which familiar participants performed with very few errors, was reflective of an impersonation scenario as lookalikes could be considered as impersonators of the celebrity. Therefore results are agreeing that familiar viewers are generally able to perform highly for matching faces that involve impersonation (Experiments 1 & 12) but evasion disguise can cause familiar viewers matching difficulty (Experiment 12).

Familiar viewers outperformed unfamiliar viewers in all tasks (including the evasion condition described above). However there are many applied situations e.g. passport security checks and monitoring of CCTV images, where it is not possible for viewers to be familiar with the faces concerned. With this in mind, I was keen to investigate methods that had previously been found to be effective in improving face-matching performance and apply them to my tasks which involved more challenging stimuli in this thesis. I found that both image manipulation (blurring of pixelated images) and crowd analysis could

improve accuracy for pixelated face-matching (Experiments 4 & 5). I also found that super-recognisers made more accurate judgments than comparison observers. This result demonstrates that super-recognisers' superior face recognition ability extends beyond good quality images and implies that their high performance does not rely solely on fine scale information in face images. This finding therefore supports the notion of recruiting super-recognisers for face recognition roles, and extends the past evidence for this proposal by showing that super-recognisers also hold a superior ability for matching very challenging image pairs as well as cooperative image pairs (e.g. White et al., 2015). Additionally, for the specific case of improving performance on images with changed camera-to-subject distance, providing accurate distance cues is a means of boosting performance (Experiment 9).

In terms of past research on deliberate disguise, my findings support the most basic finding of previous face recognition research:- disguise presence impaired face recognition. None of the present findings go against those of previous studies, however my research builds upon previous disguise research in several important ways. Past research had studied disguise with a focus solely on evasion. A key finding from the experiments presented here is that disguise cannot be understood as a unitary manipulation. I found consistent differences between evasion and impersonation disguise in terms of matching accuracy (higher accuracy for impersonation) and also ways in which the disguises were executed by the models (higher relevance on external features for impersonation). Additionally I showed that there is more to disguise than occlusion of facial features. Past research relied almost entirely on occlusion manipulations such as glasses and facial hair. Those manipulations tell us rather little about disguise. Any form of occlusion is apt to reduce matching accuracy as it reduces the available information. Moreover occlusion also would not be an effective disguise in many security scenarios, as the props that occluded features may have to be removed for identity decisions to be reached. Here I have shown that disguise can impair face recognition even without occluding features. In addition to this, I highlighted the importance of free disguise. By giving participants the freedom to disguise themselves as they wished, rather than as prescribed by the experimenters, I was able to show that people naturally use many

different methods to disguise themselves and that some methods are more effective than others. Models did far more than simply add props in creating their disguise manipulations. Investigation of what made a successful disguise showed that there was no simple disguise recipe. Contrary to the implicit assumption of previous studies, approaches to disguise are rather idiosyncratic - some disguise manipulations (e.g. manipulating hairstyle, copying or changing expression) were effective in the case of some identities but not for others. Overall, effective Evasion involved more non-feature based changes, whereas effective impersonation was created through internal and external featural manipulations.

## 7.3 Theoretical Implications

Up until now this discussion has focussed upon the practical advances gained from my research, realistic challenging images, performance enhancements and insights into deliberate disguise. The experiments that I conducted also resulted in implications for theory. Throughout the thesis I argued that my findings support a theoretical standpoint on face learning that has within person variability at its heart: what viewers are learning when they become familiar with a face is all of the different ways which that face can look (Jenkins et al., 2011). The range of possible appearance can be constructed as a form of face space that is specific for that identity. Figure 7.2 depicts this face space as a multidimensional volume, encapsulating the range of accepted appearances for that face. Any face image that falls inside the volume is accepted as that individual, and any image that falls outwith this volume is not. Exposure to a person's face helps viewers to refine the face space for that individual. Less familiar viewers also hold their face space representations of individuals but their face space is less refined as a result of their limited exposure to the face and limited experience of appearances that the face can take, making unfamiliar viewers more likely to make identity judgment errors. Face space is refined as exposure increases, therefore refinement of face space reflects the graded familiarity advantage found in Experiments 1 & 2. I found that performance on my celebrity lookalike task improved as a result of increasing familiarity of the faces concerned. For viewers who were unfamiliar with the celebrity faces, the lookalike face

image likely fell within the accepted range of faces for the celebrity leading to an incorrect identity match decision. As familiarity increased, less of these errors occurred, with the lookalike image more often falling outside the real celebrity's face space.

My findings also suggest arguments against both the configural and featural account of face recognition. Whilst configural and featural details of a face may be used to aid face recognition in some ways, I suggest that neither configural nor featural information is the key to face recognition. Experiment 7 showed that configurations of a face, as represented in images, are not constant (e.g. Kleinberg et al. 2007). In Experiment 7, distances between features underwent non-linear changes as a result of change in camera-to-subject distance. But only unfamiliar viewers were affected by the distance manipulation. Familiar viewers performed highly no matter the camera-to-subject distance, supporting the theory that familiar viewers have expertise of all the different ways that a face can look. If they have previously seen the face across a range of distance and image conditions, then photographs of the same person taken from different distances would still fall into the accepted range of faces that a familiar viewer holds for the face, but outside the accepted range of faces held by an unfamiliar viewer.

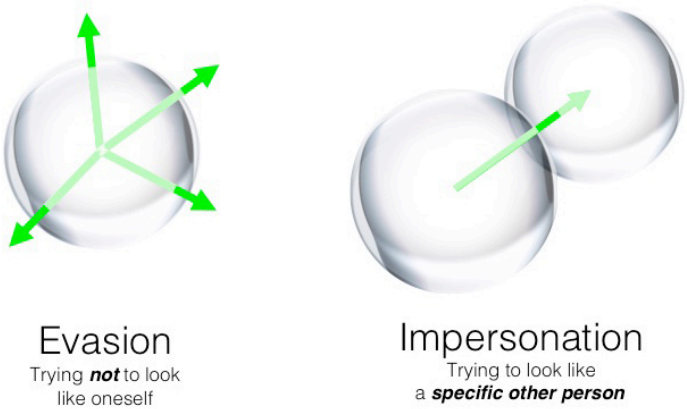


**Figure 7.1** Schematic representations of the disguise manipulations with regards to face space. Each bubble represents one individual's face space.

The theoretical standing also fits alongside the experimental results for disguise. Evasion disguise led to more errors than impersonation disguise. With relation to face space, there are more ways that a person can leave their own face space (evade identity), than enter someone else's (impersonate identity), for example changing hair colour may make someone look unlike themselves, but changing hair colour would only help for impersonation if the hair colour was changed to match that of the person that you want to look like. Impersonation involves moving appearance outwith the accepted face space for your own face, and into the accepted space of someone else. Impersonation change is thus limited to one direction (changes that take you towards, and ideally into, the face space of the person who is to be impersonated) whereas evasion can involve change in any direction as long as the face is removed out of its own face space (see Figure 7.1).

Exploration and Experiment 13 in Chapter 6 suggested a more complex component to face space with relation to disguise. Many different techniques were used by my models, and in various combinations, to create the various evasion and impersonation disguises. Some disguises worked for some identities but not so well for others. It is possible that certain faces were more unique or generic, hence making it easier or harder for the models to move out of their own face space or into another's. Making changes to a face, which lead to a change in social inference, was found to be an effective way of moving a face outwith its own face shape in the case of evasion disguise.

Evasion disguise was linked to change in social inference whereas impersonation disguise was not. Familiar viewers were impaired only by evasion of identity, not impersonation disguise. Notably this is the only disguise change that was linked to significant changes in inferred social inferences. Familiar viewers had an additional advantage for identifying impersonation faces over evasion faces. In impersonation trials the familiar viewers were familiar with both the impersonator and the person being impersonated – therefore there were two ways that they could approach the matching task. The viewer could either perceive the true identity of the impersonator, or decide that the impersonator falls outside the accepted range for the person being impersonated. In future studies it would

be possible to dissociate these possibilities by conducting a study in which viewers were familiar with only one of the faces in each pair. In any case it is clear that impersonation is harder to accomplish than evasion because the direction of change is constrained.

What is not yet known is exactly what it is that people are learning in a face when they are becoming familiar with it. Face space is proposed as a multidimensional space. An identity may change in appearance across many of these dimensions, but for recognition to occur the match must presumably fall within the expected range for at least one of the dimensions. I have talked about face space in a rather abstract representation as described by Burton et al. (2015). Dimensions that are relatively unaffected by disguise, could be identified by applying Principal Components Analysis (PCA) to multiple images (e.g. 20 undisguised images) of the model in order to establish a face space for the identity. New images of the identity could then be entered into this face space along with the disguised images to see where each of these images fall. This would show if there are any dimensions that remain constant within the disguise face images, and also perhaps reveal which dimensions of change are necessary in the creation of effective evasion and impersonation disguises. As it stands it is unclear what it is that people are recognizing when they are successfully seeing through disguise to recognizing a face. It may be that critical information is the same for every case of disguise, although it is likely to be different for evasion and impersonation. This interpretation is based on the differences found between these types of disguises so far. It is also possible that different things are used to see through each disguise created by each model because the underlying face of each of these models is different to begin with. This type of investigation for disguise could also aid the understanding of undisguised (normal) face recognition. Whatever familiar viewers can ‘see through’ in the case of disguise cannot be what is critical for identification.

## 7.4 Practical Implications

Several practical implications follow from my research. First of all there is a need to acknowledge that face-matching involving challenging images is less reliable than cooperative face-matching. Moreover, merely acknowledging this problem is not enough. In a striking demonstration of this (Experiment 11), being aware that a face may be disguised did not improve matching performance. Instead specific methods such as blurring of pixelated images (Experiment 4), crowd analysis of identity judgments (Experiment 5) and personnel selection (Experiment 6) will likely be required. Even then errors will not be eliminated entirely.

Of all the manipulations in this thesis, familiarity improved performance most. Familiar viewers are thus recommended to make identity judgments for similar challenging images in forensic situations. As the graded effects in Experiments 1-4 emphasise, a little familiarity with a face is better than none at all, and a lot is better than a little. In summary, the more familiar a person is with the face or faces concerned, the more likely the correct identity judgment will be reached.

In Experiment 8, I found that photographic conditions – specifically camera-to-subject distance – could impair identity judgments for unfamiliar faces. My Experiments 2 & 3 show a similar decrease in performance for pixelated faces. However in Experiment 8, I confirmed that changing the distance from which a photograph is taken from can make it difficult to compare identity across images. This finding highlights that using photographs for identity comparisons is problematic even when image quality is high. Distance cues can aid recognition when these are available, but cues are not readily available from most photograph images, especially images used for identity confirmation, as these images are normally cropped around the face, removing any background related cues that may have indicated subject to camera distance. I found that images are more accurately matched when taken from the same distance. A practical implication from this would be that where possible, consistency should be applied when photographing individuals in security

and forensic scenarios. For example, police could photograph all suspects from the same distance across all photos of an individual. This standardisation would make it easier to identify the same person across multiple images and also make it less likely that a suspect would be mistaken for another person in a side-by-side image comparison scenario. Camera-to-subject distance changes would however remain problematic for archived images and some CCTV footage, but introducing standardisation to station captured images may aid some identity scenarios. Using familiar viewers to identify faces taken of different distances will largely overcome the problem as they are far less affected by camera-to-subject distance change.

My disguise research has implications for applied settings. I show that people are very poor at matching disguised face images, suggesting that it is rather easy to carry out a successful disguise. Disguises that involve evading identity are more likely to attract errors than disguises that involve the impersonation of somebody else. There are practical implications here in terms of both criminal cases of identity fraud and also undercover policing. In terms of criminal disguise activity, my research highlights that cases of impersonation and evasion may go undetected. Successful evasion disguise is particularly concerning as disguise could make it particularly difficult to catch a suspect. Impersonation disguise carries many security threats in terms of identity fraud, which could have catastrophic results particularly in terms of allowing a dangerous or unauthorised person into a country or providing them with access to information, which their true identity should not be granted. Disguise is problematic for face recognition, and errors can pose danger. Therefore steps discussed above should be taken to try and minimise fraud and identity evasion through disguise. The disguise results do however suggest that undercover police can successfully keep their true identity hidden by using evasion disguise.

There are also some important implications from this thesis in relation to the experimental psychology research practice for the study of face recognition. First my findings reiterate the importance of a distinction between familiar and unfamiliar face-

matching, and extend this to provide further evidence for a graded familiarity advantage in face recognition. Past studies have tested face-matching performance using cooperative face-matching stimuli. I suggest that measuring performance for cooperative stimuli images does not capture face-matching performance levels for uncooperative stimuli. I quantified the performance costs for various challenging situations and found lower accuracy levels for the images that I tested, than those reported by previous studies. It is important to study uncooperative image performance, especially when face recognition performance is so closely linked to many security systems or included in witness testimonials for crimes. Suspects may deliberately make their identification effort difficult. I also highlight important distinctions within disguise. The previous research on disguise was based on evasion. I have extended upon this in two ways; i) by investigating evasion disguise as being more than just the addition of prescribed props which occlude facial features, and ii) by looking at impersonation which had been completely ignored by past studies of disguise. Individuals spontaneously use many different types of manipulations to disguise themselves, and these disguises impair identity performance even when there is no occlusion of internal facial features. My research has found clear distinctions between evasion and impersonation, both in terms of face-matching performance and the kinds of disguise manipulations applied. Future research will have to acknowledge these distinctions if we are to arrive at a complete understanding of deliberate disguise.

## 7.5 Future Directions

This thesis expands the research for face recognition in challenging situations, however there are areas of investigation that I believe warrant future study. Firstly, it would be interesting to test face memory performance for the FAÇADE database images. I found poorer performance accuracy for matching disguised faces than undisguised faces. Matching faces is presumably an easier task than face recognition tasks that involve a memory component (Megreya & Burton, 2008). Previous memory studies have tested for recognition of disguised faces in direct comparison to performance for memory of identical images (Patterson & Baddeley, 1977). My FAÇADE database includes different

images of the same faces undisguised and in disguise, and evasion and impersonation disguise images. This database would thus allow a more thorough and experimentally sound test of face memory performance for disguised versus undisguised faces, and also allow performance for evasion disguise to be compared with impersonation disguise. Performance could be even worse for disguise in a memory task, as in the matching task both images could be compared in side-by-side comparison. It is not obvious whether the differences in performance found for evasion and impersonation would remain in a memory scenario. I believe this to a be a particularly interesting comparison, as impersonation could trick viewers to incorrectly 'remember' the real target face and reject an image presented of the true identity undisguised. There are many reports of erroneous eyewitness memory for studies that do not include impersonation disguise (Bruce, 1988), therefore impersonation disguise may increase error rates. Additionally, evasion disguise may result in rejection of the true identity face presented in an undisguised form.

It would also be interesting to test whether super-recognisers may also be able to improve performance accuracy in the case of disguise images. This would need to be confirmed by future research. If super-recognisers do perform better than our unfamiliar, or even possibly familiar controls, then super-recognisers could be of great help in identity efforts involving disguise. It would be important to explore whether super-recognisers or familiar viewers provide a more accurate viewer group to call on for making identity decisions. Based on the only directly comparable data in this thesis for an unfamiliar group of super-recognisers (Experiment 6 [mean performance accuracy for this group = 76%]) and the highest familiarity group of comparison participants (Experiment 2 [mean performance accuracy for this group = 76%]), I would predict that unfamiliar super-recognisers and familiar comparison viewers are equally good at matching faces.

Another interesting investigation for the future research would be to explore the effect of methods of improving performance (Experiments 4, 5 and 6) in the combinations that were not addressed in this thesis. I found that when I tested methods in combination, this

led to further improvements in performance than any of the methods used alone. Due to the limited opportunity to work with super-recognisers I was unable to test all of these combinations. Specifically, it is currently unknown whether blurring the pixelated images would have further improved the performance of super-recognisers, and also whether combining all three methods could result in improvements greater than those from combining any two methods.

A key future experiment will be the Principal Component Analysis (PCA) of disguised faces outlined in the theoretical advances section in this discussion. The proposed experiment will allow theoretical advancement in terms of what it is that allows viewers to see through disguise. This has particular relevance to the theory of face space, and the refinement of face space with familiarity. The familiarity component of this theory can be explored through manipulating the number of reference face images entered into the PCA. This will help to establish what it is about a face which stays constant in disguise, and whether there are reliably constant factors for it at all, or if factors differ by identity.

Finally, I believe it would be interesting to investigate machine performance for disguised faces. Computer algorithms are now being used in many security situations including passport security, but as with human performance, machine performance has generally been tested using cooperative stimuli. Attempts to test machine performance for disguised faces have many of the same limitations as previous human disguise investigations, namely testing has been limited to evasion disguise where images were disguised through occlusion of features. The props which cause occlusion disguise would be removed in most security scenarios, therefore testing machine performance on these images is of limited use for security algorithms used in passport control. Real cases of identity fraud more often include impersonation or the morphing of a new holder's face with the face of the true passport owner. Impersonation is therefore particularly relevant in terms of security related identity fraud, but has not yet been tested in terms of machine performance. My FAÇADE database provides stimuli void of occlusion of features, and includes evasion and impersonation disguise. Future studies could therefore

use this database to more fully explore machine performance for disguised faces and compare this performance for undisguised versions of the same faces. As data has already been collected for both unfamiliar and familiar human viewers, machine performance could then be directly compared to performance for each of these human viewer groups, helping to answer the question of whether humans or machine are better at matching disguised faces. Greater performance accuracy could perhaps be achieved through fusing the performance of humans and machine, this has been a successful method of improving face recognition accuracy in past studies (O'Toole et al. 2007).

In summary, previous estimates of identification accuracy are likely to be overestimates if the people who were being identified were cooperating with their identity effort. Applied cases of face recognition will more likely include images with incidental or deliberate differences to their own appearance or similarities with another person, as have been investigated by this thesis. Familiar viewers are generally considered to be exceptionally good at identifying faces and this thesis has shown that this familiarity advantage indeed extends to performance for both incidental and deliberate appearance change. However, challenging situations start to reveal limits of this advantage. For example, familiar viewers' face-matching performance is impaired by evasion disguise – this raises interesting questions about what appearance change a familiar viewer will allow to classify an individual as still being the same person. The results of this thesis suggest that familiar viewers are better than unfamiliar viewers for the experience of a face that they have learnt. There are also many applied implications of this research. For example, identity situations in the real world are often linked to impersonation disguise, however there has been a great deal of disconnect between lab experiments and the real world problem of identity fraud, as impersonation had not been addressed by previous lab experiments. I found that impersonation is harder to pull off than evasion disguise. Poor performance for matching evasion disguise faces is problematic for capturing criminals or missing people on police watch lists, but suggests that evading identity of undercover police officers will likely be effective. There is still a lot more to learn about face recognition in challenging situations, especially for the case of disguise, and I hope the FAÇADE database will continue to aid this investigation.

# References


Alain, C., & Proteau, L. (1980). Decision making in sport. In C.H. Nadeau, K.M. Halliwell, K.M. Newell, & G.C. Roberts (Eds.), Psychology of motor behaviour and sport. 465-477. *Champaign, IL*: Human Kinetics.

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage*, *23*(3), 905–913. doi:10.1016/j.neuroimage.2004.07.060

Aristotle, Jowett, B., & Davis, H. W. C. (1920). Aristotle's Politics. Oxford: At the Clarendon Press.

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3,* 85-103.

Behrmann M, Avidan G 2005 Congenital prosopagnosia: face-blind from birth. *Trends Cogn. Sci*. **9**, 180–187

Bindemann, M., Attard, J., Leach, A. M. Y., & Johnston, R. A. (2013). The Effect of Image Pixelation on Unfamiliar-Face Matching, *717*(November), 707–717.

Bindemann, M., Burton, A.M., Leuthold, H., & Schweinberger, S.R. (2008). Brain potential correlates of face recognition: Geometric distortions and the N250r brain response to stimulus repetitions. *Pscyhophysiology*,*45*(4), 535-544.

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS ONE*, *11*(2), 1–13. doi:10.1371/journal.pone.0148148

Bonner, L., Burton, A.M. & Bruce, V. (2003). Getting to know you: how we learn new faces. *Visual Cognition*. *10*, 527-536.

Borkenau, P. Brecke, S. Mottig, C., & Paelaecke, M. (2009). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*, 645-657.

Bradshaw, J.L., & Wallace, G. (1971). Models for the processing and identification of faes. *Perception and Psychophysics*, *9*, 443-448.

Brigham, J.C.& Bothwell, R.K. (1983). The ability of prospective juroros to estimate the accuracy of eyewitness identification. *Law and Human Behaviour*. *7*, 19-30.

Bruce, V. (1982). Changing faces: visual and non-visual coding in face recognition. *British Journal of Psychology*. *73*, 105-116.

Bruce, V. (1986). Influences of familiarity on the processing of faces. *Perception*, *15*(4), 387–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3822723

Bruce, V. (1988) *Recognising Faces*. London: Erlbaum.

Bruce, V. (1994). Stability from variation: The case of face recognition the M.D. Vernon memorial lecture. The Quarterly Journey of Experimental Psychology Section A: Human Experimental Psychology/abstract content. *47(1)*, 5-28.

Bruce, V., Doyle, T., Dench, N. & Burton, M. (1991). Remembering facial configurations. *Cognition, 38(2*), 109-144.

Bruce, V., Henderson, Z, Greenwood, K., Hancock, P.J.B. & Burton, A.M. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology – Applied*, *5*, 339-360.

Bruce, V., Henderson, Z., Newman, C., Burton, A.M. (2001). Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV. *Journal of Experimental Psychology: Applied, 7*, 207-218.

Bruce, V. & Langton, S. (1994). The use of pigmentation and shading information in recognising the sex and identity of faces. *Perception, 23*, 803-822.

Bruyer, R. & Coget, M. C. (1987). Features of laterally displayed faces: saliency or top down processing? *Acta Psychologica*, *66*, 103-114.

Bryan, R., Perona, P. & Adolphs, R. (2012). Perspective Distortions from Interpersonal Distance Is an Implicit Visual Cue for Social Judgments of Faces. PLoS ONE 7(9): e45301. doi:10.1371/journal.pone.0045301

Burton, a. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2015). Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cognitive Science*, *40*, 1–22. doi:10.1111/cogs.12231

Burton, a. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments Against a Configural Processing Account of Familiar Face Recognition. *Perspectives on Psychological Science*, *10*(4), 482–496. doi:10.1177/1745691615583129

Burton, a. M., & Vokey, J. R. (1998). The Face-Space Typicality Paradox: Understanding the Face-Space Metaphor. *The Quarterly Journal of Experimental Psychology Section A*, *51*(3), 475–483. doi:10.1080/713755768

Burton, A. M., White, D., & McNeil, A. (2010). The Glasgow Face-matching Test. *Behaviour Research Methods*, *42*(1), 286-291.

Burton, A. M., Wilson, S., Cowan, M., Bruce, V. (1999) Face recognition in poor-quality video: Evidence from security surveillance. Psychol Sci 10: 243–248. doi: 10.1111/1467-9280.00144

Clutterbuck, R., & Johnston, R. a. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985–994. doi:10.1068/p3335

Clutterbuck, R., & Johnston, R. a. (2004). Demonstrating the acquired familiarity of faces by using a gender-decision task. *Perception*, *33*(2), 159–168.

Clutterbuck, R., & Johnston, R. a. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116. doi:10.1080/09541440340000439

Coates, Tim (1999). The strange story of Adolph Beck. *London*: Stationery Office.

Collishaw, S. M., & Hole, G. J. (2000). Featural and configurational processes in the recognition of faces of different familiarity. *Perception*, *29*(8), 893–909. doi:10.1068/p2949

Costen, N. P., Parker, D. M.m & Craw, I. (1994). Spatial content and spatial quantization effects in face recognition. *Perception, 23,* 129-146.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception and Psychophysics, 58,* 602-612.

Davies, G., Ellis, H. D., & Shepherd, J. (1978). Face recognition accuracy as a function of mode of representation. *Journal of Applied Psychology*, *63*(2), 180–187. doi:10.1037//0021-9010.63.2.180

Dhamecha, T. I., Singh, R., Vasta, M. & Kumar, A. (2014). Recognising disguised faces: Human and machine evaluation. *PlosOne*.*9*(7): e99212.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General. 115*, 107-117.

Donders, F. C. (1868/1969). On Speed of Mental Processes. *Acta Psychologica, 30*, 412-431.

Dowsett, A. J., & Burton, a. M. (2014). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, n/a–n/a. doi:10.1111/bjop.12103

Duchaine, B. (2011). Developmental Prosopagnosia: Cognitive, Neural and Developmental Investigations. In A. J. Calder., G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford Handbook of Face Perception* (pp.821-838). Oxford, UK: Oxford University Press.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face

stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. doi:10.1016/j.neuropsychologia.2005.07.001

Ellis, H.D., Shepherd, J.W. & Davies, G.M. (1979). Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception*, *8*, 431-439.

Flow, H & Ebbesen, B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. *Law and Human Behaviour*. 31(1), 33-52.

Galper, R.E. (1970) Recognition of faces in photographic negative. *Psychonomic Science*,*19*, 207-208.

Galton, F. (1907). Vox populi - The wisdom of crowds. *Nature*, *75*, 450–451. doi:10.1038/075450a0

Ghent, L. (1960). Recognition by children of realistic figures presented in various orientations. *Canadian Journal of Psychology*, *14*, 249-256.

Golstein, A.G. (1965). Learning of inverted and normally orientated faces in children and adults. *Pscyhonomic Science*, *3*, 447-448.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203. doi:10.1016/S0896-6273(00)80832-6

Haig, N. D. (1984). Faces in Perception. *Perception*, *13*, 1158–1165.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces, *4*(9), 330–337.

Harmon, L. & Julesz, B. (1973). Masking in Visual Recognition: Effects of Two-Dimensional Filtered Noise. *Science. New Series, 180.* 1194-1197.

Harper, B. & Latto, R. (2001). Cyclopean vision, size estimation, and presence in orthostereoscopic images. *Presebce*, *10*(3), 312-330.

Henderson, Z., Bruce, V., & Burton, a. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*(4), 445–464. doi:10.1002/acp.718

Henle, M. (1942). An experimental investigation of past experience as a determinant of visual perception. *Journal of Experimental Psychology*, *30*, 1-22.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(4), 986. doi:10.1037/0096-1523.22.4.986

Hochberg, J. & Galper, R.E. (1967). Recognition of faces: An exploratory study. *Psychonomic Science*, *9*, 619-620.

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*. doi:10.1068/p3252

Home Office. (2012). False ID Guidance. *The Home Office.* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/98108/false-id-guidance.pdf

Huff, CR. (1987). Wrongful Conviction: Societal tolerance of injustice. *Research in social problems and public policy*. *4*, 99-115.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1571), 1671–1683. doi:10.1098/rstb.2010.0379

Jenkins, R., & Kerr, C. (2013). Identifiable images of bystanders extracted from corneal reflections. *PLoS ONE*, *8*(12), 8–12. doi:10.1371/journal.pone.0083325

Jenkins, R., White, D., Van Montfort, X., & Burton, A.M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–23. doi:10.1016/j.cognition.2011.08.001

Johnston, A. Hill, H. & Carman, N. (1992). Recognising faces: effects of lighting direction, inversion and brightness reveral. Perception, 21, 365-375.

Kanade, T. (1977). Computer recognition of human faces. *Stuttgart*: Birkhauser Verlag.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761. doi:10.1016/S0896-6273(00)80734-5

Kelly, M.D. (1970). Visual identification of people by computer. Tech. rep. AI-130, Stanford AI Project, Stanford, CA.

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

Kemp, R., Pike, G., White, P., & Musselman, a. (1996). Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues. *Perception*, *25*(1), 37–52.

Kenny, D. a., Horner, C., Kashy, D. a, & Chu, L. C. (1992). Consensus at zero acquaintance: replication, behavioral cues, and stability. *Journal of Personality and Social Psychology*, *62*(1), 88–97. doi:10.1037/0022-3514.62.1.88

King, a. J., Cheng, L., Starke, S. D., & Myatt, J. P. (2012). Is the true “wisdom of the crowd” to copy successful individuals? *Biology Letters*, *8*(2), 197–200. doi:10.1098/rsbl.2011.0795

Kleinberg, K. F., Vanezis, P., & Burton, a M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences*, *52*(4), 779–83. doi:10.1111/j.1556-4029.2007.00458.x

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: Diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948. doi:10.1016/j.anbehav.2010.12.018

Lander, K., Bruce, V., & Hill, H. (2001). Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology, 15,* 101-116.

Leder, H., & Carbon, C.C. (2006). Face-specific configural processing of relational information. *British Journal of Psychology, 97*, 19-29.

Leikas, S., Verkasalo, M., & Lönnqvist, J. E. (2013). Posing personality: Is it possible to enact the Big Five traits in photographs? *Journal of Research in Personality*, *47*(1), 15–21. doi:10.1016/j.jrp.2012.10.012

Light, L.L.; Kayra-Stuart, F. & Hollander, S. (1979) Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory.* Vol 5(3), 212-228. http://dx.doi.org/10.1037/0278-7393.5.3.212

Liu, C.H. (2003). Is Face Recognition in Pictures Affected by the Center of Projection? *IEE International Workshop on Analysis and Modeling of Faces and Gestures*, (Los Alamitos, CA: IEEE Computer Society), 53-59.

Loftus, E. & Doyle, J. (1992). Eyewitness testimony: Civil and Criminal. *Charlottesville, VA*, Lexis Law.

Matinez, A.M & Benavente, R (1998). The AR Face Database. *CVC Technical Report.* Number 24.

Maurer, D., Grand, R. Le, & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12039607

Megreya A.M., Burton A.M. (2007) Hits and false positives in face-matching: A familiarity-based dissociation. *Perception & Psychophysics*. *69*, 1175–1184. doi: 10.3758/bf03193954

Megreya, A. M., Burton, a. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34* (4), 865–876.

Megreya, A.M., & Burton, a.M. (2008). Matching faces to photographs. Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14(4),* 364-372.

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

Menon, A., White, D. & Kemp. R. (2015). Variation in Photos of the Same Face Drives Improvements in Identity Verification. *Perception*. doi: 10.1177/0301006615599902

Morrone, M, C., Burr, D, C., Ross, J. (1983). Added noise restores recognisability of coarse quantized images. *Nature, 305,* 226-228.

O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, *30*(6), 755–764. doi:10.1068/p3027

O'Toole, A.J., Edelman, S. & Bulthoff, H.H. (1998). Stimulus specific effects in face recognition over changes in viewpoint. *Vision Research*. 38, 2351-2363.

Patterson, K.E. & Baddeley, A.D. (1977). When Face Recognition Fails. *Journal of Experimental Psyhcology: Human Learning and Memory*, *3*(4), 406-417.

Pirenne, M. H. (1970) Optics, Painting, & Photography. *New York*: Cambridge University Press.

Pizlo, Z. (1994). A theory of shape constancy based on perspective invarients. *Vision Research*. 34(12). 1637-1658.

Ramanathan, N., Chowdhury, A.R. & Chellappa, R. (2004). Facial Similarity Across Age, Disguise, Illumination and Pose. *Proceedings of International Conference on Image Processing*, 3, 1999-2002.

Rhodes, G., Brennan, S. & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology. 19*, 473-497.

Righi, G., Peissig, J.J. & Tarr, M.J. (2002). Recognising disguised faces. *Visual Cognition*. *20*(2), 143-169.

Robbins, R., & Mckone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition, 88*, 79–107. doi:10.1016/S0

Robertson, D. J., Noyes, E., Dowsett, A., Jenkins, R., Burton, A. M., & Burton, M. (2016). Face recognition by Metropolitan Police Super-recognisers. *PLoS ONE, 11(2):* e0150036. doi:10.1371/journal.pone.0150036

Rossion, B. (2009). Distinguishing the cause and consequence of face inversion: the perceptual field hypothesis. *Acta Psychologica*, *132*(3), 300–12. doi:10.1016/j.actpsy.2009.08.002

Rule, N. O., & Ambady, N. (2011). Judgments of Power From College Yearbook Photos and Later Career Success. *Social Psychological and Personality Science*, *2*(2), 154–158. doi:10.1177/1948550610385473

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: people with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. doi:10.3758/PBR.16.2.252

Sandford, A., & Burton, a. M. (2014). Tolerance for distorted faces: Challenges to a configural processing account of familiar face recognition. *Cognition*, *132*(3), 262–268. doi:10.1016/j.cognition.2014.04.005

Sergent, J. (1986). Microgenesis of face perception. In *Aspects of Face Processing* (Ellis, H. D. et al., eds). 17-33, Martinus Nijhoff.

Singh, R., Vatsa, M. & Noore, A. (2009). Face Recognition with Disguise and Single Gallery Images. *Image and Vision Computing,* 27(3), 245-257.

Smith, E.E. & Nielsen, G.D. (1970). Representations and retrieval processes in short-term memory: Recognition and recall of faces. *Journal of Experimental Psychology*, *85*, 397-405.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology*, *89*(1), 39–57. doi:10.1111/j.2044-8295.1998.tb02672.x

Sutherland, C. a M., Oldmeadow, J. a, Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–18. doi:10.1016/j.cognition.2012.12.001

Tanaka, J. W. & Farah, M.J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*, 225-245.

Terry, R.L. (1993). How wearing eye glasses effects facial recognition. *Current Psychology*. *12*(2). 151-162. 483-492.

Terry, R.L. (1994). Effects of facial transformations on accuracy of face recognition. *The Journal of Social Psychology*. 134(4), 151-162.

Thomson, D.M. (1986). Face recognition: More than a feeling of familiarity? In H.D. Ellis, M.A. Jeeves, F. Newcombe & A. Young (Eds.), *Aspects of face processing*. Dordrecht: Matinus Nijhoff.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions, (October), 455–460. doi:10.1016/j.tics.2008.10.001

Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The significance of hair for face recognition. *PLoS ONE*, *7*(3), 1–8. doi:10.1371/journal.pone.0034144

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2-3), 214–218. doi:10.1068/p7676

Troje, N.F. & Bulthoff, H.H. (1996). Face recognition under varying posers: the role of

texture and shape. *Vision Research*, 36, 1761 -1771.

Uttal, W, R., Baruch, T., Allen, L. (1995). Sequential image degradations in a recognition task. *Perception & Psychophysics, 57,* 682-691.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. Quarterly Journal of Experimental Psychology Section A, 43(2), 161-204.

Webster,M. & Mollen, J. (1995). Colour constancy influenced by contrast adaptation. *Nature*. 373, 694-698.

Wells, G., Rydell, S. & Seelau, E. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology.* 78(5). 835-844.

Wells, G.L., Small, M., Penrod, S.D, Malpass, R.S., Fulero, S.M., Brimacombe, C.A.E. (1998). Eyewitness identification procedures: recommendations for lineups and photspreads. *Law and Human Behaviour*. *22*, 603-607.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. O. B. (2013). Crowd Effects in Unfamiliar Face Matching, *777*(November), 769–777.

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*(1), 100–6. doi:10.3758/s13423-013-0475-3

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face-matching. *PloS One*, *9*(8), e103510. doi:10.1371/journal.pone.0103510

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE*, *10*(10), 1–14. doi:10.1371/journal.pone.0139827

White, D., Phillips, P. J., Hahn, C. A., Hill, M., Toole, A. J. O., & White, D. (2015b). Perceptual expertise in forensic facial image comparison, Proceedings of the Royal Society, 282, 1–8.

Winograd, E. (1976). Recognition memory for faces following nine different judgments. *Bulletin of the Pscyhonomic Society, 8*, 419-421.

Woodhead, M. M., Baddeley, a. D., & Simmonds, D. C. V. (1979). On Training People to Recognize Faces. *Ergonomics*, *22*(3), 333–343. doi:10.1080/00140137908924617

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145.

Young, A.W., Hay, D.C., McWeeny, K.H., Flude, B.M. & Ellis, A.W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*, 737-746.