Facial creation: using compositing to conceal identity

Sarah Louise Shrimpton

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with the University of Central Lancashire, School of Psychology

January 2018

Abstract

This study focused on the creation of new faces by compositing features from donor face photographs together that provide a way to generate new face identities. However, does the act of compositing conceal the identity of the donor faces? Two applications of these created faces require donor face identities to remain concealed: Covert social media profiles provide a way for investigating authorities to survey online criminal activity and, as such, a false online identity, including face image, is required. Compositing features/face parts from various donor face photographs could be used to generate new face identities. Face donor photographs are also used for the 'texturing' of facial depictions to reconstruct an image of how a person might appear. This study investigated whether compositing unknown face features onto known familiar faces (celebrities and lecturers) was sufficient to conceal identity in a face recognition task paradigm. A first experiment manipulated individual features to establish a feature saliency hierarchy. The results of this informed the order of feature replacement for the second experiment, where features were replaced in a compound manner to establish how much of a face needs to be replaced to conceal identity. In line with previous literature, the eyes and hair were found to be highly salient, with the eyebrows and nose the least. As expected, the more features that are replaced, the less likely the face was to be recognised. A theoretical criterion point from old to new identity was found for the combined data (celebrity and lecturer) where replacing at least two features resulted in a significant decrease in recognition. Which feature was being replaced was found to have more of an effect during the middle part of feature replacement, around the criterion point, where the eyes were more important to be replaced than the mouth. Celebrities represented a higher level of familiarity and, therefore, may be a more stringent set of results for practical use, but with less power than the combined data to detect changes. This would suggest that at least three features (half the face) need to be replaced before recognition significantly decreases, especially if this includes the more salient features in the upper half of the face. However, once all six features were replaced, identity was not concealed 100% of the time, signifying that feature replacement alone was not sufficient to conceal identity. It is completely possible that residual configural and contrast

information was facilitating recognition, and, therefore, it is likely that manipulations, such as these, are also needed in order to conceal identity.

Contents

List of Figures

List of Tables

Contributors and Acknowledgements

First, I want to thank my Mum, Dad and sister; their unwavering support has given me the will and courage to complete this project. When doubt crept in, they were always there to buoy me up and reassure me. I am so honoured to know what it feels like to be unconditionally loved. Further thanks goes to the rest of my family and friends for their continued encouragement and understanding.

I would like to thank my supervisors, Charlie Frowd, Chris Rowland and Caroline Wilkinson, whose guidance, inspiration and expertise has been invaluable. I am indebted to Charlie's patience and wisdom, without which I would not have completed this thesis, and to Caroline, whose perseverance and energy got this project off of the ground.

My thanks also to the team in Face Lab and colleagues in Dundee whose support has helped me in so many ways, and to Stenton Mackenzie, whose continued understanding and emotional guidance has been important to my well-being.

I am grateful to the participants that donated their time to make this research possible and to the efforts of staff at Liverpool John Moores University, the University of Central Lancashire and the University of Dundee.

This research was supported by funding from the Defence, Science and Technology Laboratory (awarded to Professor Caroline Wilkinson) and carried out in collaboration with the University of Central Lancashire.

Authors Declaration

I declare that no material contained in the thesis has been used in any other submission for an academic award.

1. General Introduction

1.1 The importance of faces

Each face is unique (Lucas and Henneberg, 2015). We use faces every day to communicate with each other, through expressing emotion and speech, as well as to identify each other (Bruce and Young, 1986), relying on the seemingly fine differences between each face that make them different from one another (Little et al., 2011). Recognition of familiar faces is a remarkable feat considering faces make for a very homogenous group made up of the same component parts; two eyes above a nose, above a mouth surrounded by a facial contour and flanked by two ears. It is generally thought that a combination of different streams of information gleaned from the face allow us to differentiate between them, such as the textural pattern (e.g. dark pupils against a white sclera), featural detail, such as the shape of the mouth, as well as the spatial layout of those features (Rhodes, 1988, Cabeza and Kato, 2000, Gilad-Gutnick et al., 2012, Maurer et al., 2002, Sergent, 1984) although their relative importance has been contested more recently (Burton et al., 2015). This heightened attention to faces also fuels the phenomenon called, Face Pareidolia - seeing face-like patterns in everyday objects/scenes (Liu et al., 2014), suggesting we may be programmed to look for faces. Faces are considered so important that society deems them a key tool for identifying someone in crucially important scenarios, such as passport control or for ID cards, eyewitness testimony (del Carmen and Walker, 2012) and even for automated face recognition systems (Zhao et al., 2003). With these systems, faces can be used as a biometric in high security environments such as airports and buildings with restricted access (Parmar and Mehta, 2013, Sanchez del Rio et al., 2016) and for investigative procedures such as searching for criminals (Klontz and Jain, 2013). Our faces, in addition to identity, provide a plethora of information for social interaction and communication through expression (Ekman, 2006) and speech as well as to “house” our main senses – sight, hearing, taste, touch and smell. Perhaps because of this requirement for social structure and interaction, our faces have evolved to be unique so that we can present an individual identity within those social structures and because we lacked, or lost, the sensitivity to methods used by other animals, such as identity specific smell or vocalization (Sheehan and Nachman, 2014).

In the contemporary world, our faces have become muses, as well as gatekeepers, within the digital realm through the advent of digital technology and the world-wide web. Humans are fascinated with digital representations and documentations of faces, with the more recent obsession with the so-called, 'Selfie' (Souza et al., 2015) and the use of computer generated imagery (CGI) for highly photo-realistic digital faces (Tiddeman, 2012). As such, the development of technology has allowed for the creation of new face images, which can behave as either a new face or a proxy for existing ones.

1.2 Facial creation

Created faces allow for the freedom to generate an image that can be a completely new identity, or an adaptation or representation of an existing one. There is a range of quality of created face images from animated cartoon type representations (e.g. avatars representing players in basic video games or testing paradigms) (Rhee and Lee, 2013) to some applications allowing for photo-realistic created faces either through the superimposition of an existing real face photograph/3D scan (Aitpayev and Gaber, 2012, Lyons et al., 1998) or through the digital generation of a (often photo-realistic) face image (CGI) (Chai et al., 2003, Blanz and Vetter, 1999). There are various digital applications that utilise created faces, both offline and online, from computer gaming, training packages, learning environments, police investigations, clinical care, psychological studies to communication applications and social media platforms, to name but a few (Cosatto et al., 2003, Xie et al., 2015, Gaggioli et al., 2003, Segovia et al., 2012).The focus of this postgraduate research, was the use of created faces in forensic contexts.

Forensic applications of created faces

Created faces have been used in a variety of forensic scenarios. It is important that faces created for forensic purposes (and in some cases for other scenarios too), make sure to not only represent a new identity, but conceal the identity of any constituent face parts that have been used to create the new face. This study reports on two different forensic scenarios that make use of created faces:

1. Created face avatars[1] for false (covert) online social media profiles, and
2. Facial depictions used to prompt a potential identification in investigations.

1. Social media, in particular, has perpetuated online networking across many different genres and as such has seen a dramatic rise in the use of social media 'profiles' and chat rooms for engaging in networking for criminal activity (Europol, 2017, Martellozzo, 2013). In these particular instances a fake profile and face image is generated, called an avatar. The use of created faces that behave as avatars has become prevalent in these contexts due to the opportunity to protect the real criminal identity by allowing them to provide a fake profile photo (created avatar) and pseudonym, whilst still engaging and interacting with fellow users (often via chat-rooms): by using a created face representation, there is no need to use other veridical images, such as a personal photograph or video, that would reveal their identity, or, run the risk of using someone else's face image that may indicate to fellow users and investigators that the profile is suspicious. This method of interaction, networking and dissemination is, in particular, prevalent in online child sexual exploitation where offenders use social networking sites and chat rooms to groom children as well as interact with fellow offenders and to disseminate child abuse material (CEOP, 2013, Europol, 2017, UNODC, 2014).

Within the UK, the Metropolitan Police have their own Paedophile and High Tech Crime Unit, part of the Sexual Offences Command, who tackle the very issue of online grooming and sharing of indecent images of sexual exploitation and abuse of children that use social media platforms for networking (Gallagher, 2015). In order to combat this type of

[1] The word, 'Avatar' comes from the concept of 'Descent' in Hinduism and refers to the incarnation of a deity. LOCHTEFELD, J. G. 2002. *The Illustrated Encyclopedia of Hinduism: A-M*, Rosen.

activity, investigating authorities need to infiltrate these networks for covert surveillance as well as to potentially communicate with the individuals involved to gain access to information that may be used to identify and convict them. To facilitate this process, a fake online identity, like the ones the offenders use, is usually created for use on social networking platforms. The identity needs to be convincing as that of a real person and, again, requires a fake face profile image in order to do so, but the question arises of where fake face profile images can be acquired? There are various ways of creating a new fake face image (software packages and online facilities) that could potentially be used in these contexts. The Dutch branch of Terres des Hommes, a children's rights charity, collaborated with digital animators, Lemz, to create a photo-realistic 3D animated avatar, named Sweetie, to be used specifically for a false online identity for the covert surveillance of the online sexual exploitation of children (Hommes, 2013a) (see Figure 1.2-1).

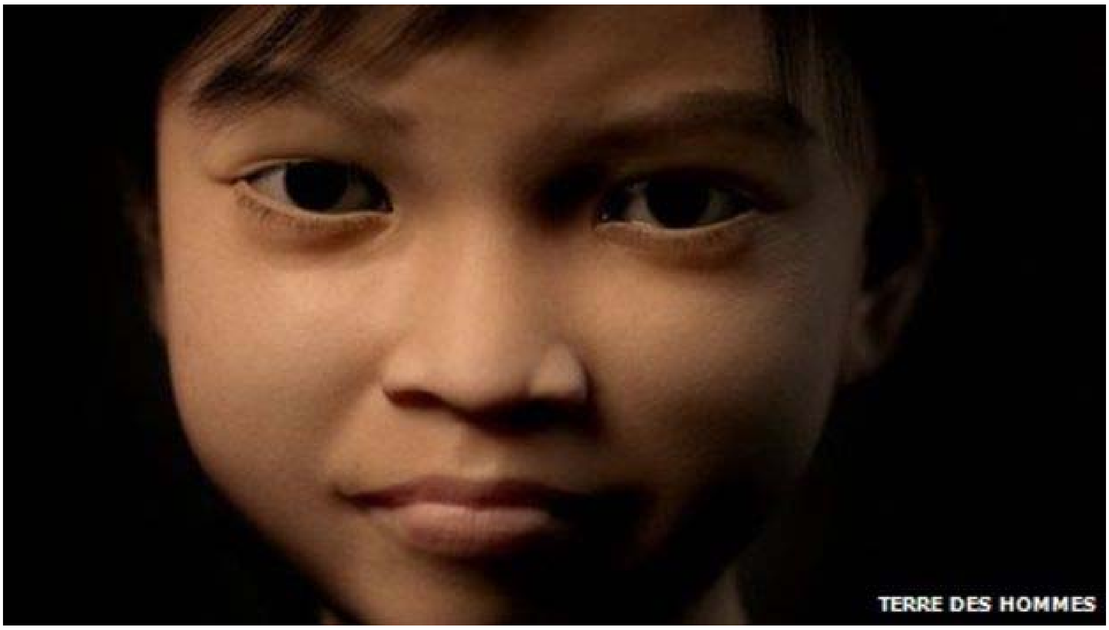


Figure 1.2-1: Sweetie, the 3D animated avatar generated by Terre des Hommes

Sweetie was used to create a false online identity for the covert surveillance of the online sexual exploitation of children. Image courtesy of Terres des Hommes.

Sweetie was animated in real-time and used for a covert sting operation involving Webcam Child Sex Tourism where adults view streaming videos of children performing sex acts (Hommes, 2013b). However, high quality 3D CGI animated faces are costly and take considerable time to generate and once Sweetie had been used in this sting operation, she was retired. Therefore, it seems important that a faster technique of generating multiple new identities (adults and children) that pass as real people for the covert surveillance of any online criminal activity is developed. This PhD study proposes a methodology where donor features/parts (e.g. eyes, nose, mouth etc.) are sampled from existing 2D face databases (photographs) (including the Glasgow face database (Burton et al., 2010)) and composited to form new faces (unknown identities). However, it is not known if the act of compositing conceals the identity of the donors within the new whole face context, which is important if the features/parts are sampled from the faces of individuals who do not wish to be recognised (e.g. investigating officers' photographs are likely to be available to the organisations creating these types of images).

2. The second scenario is that of forensic facial depictions, where created faces are used in forensic identification investigations. In these scenarios facial depictions (often created by compositing sampled features/parts from existing faces) can help promote the identification of both the living and the dead by behaving as a digital depiction of how the person might look. One example of when compositing for forensic facial depictions fails, is that of an age-progression produced by the United States' Federal Bureau of Investigation (FBI).They created an image of the al-Qaeda leader, Osama Bin Laden, in 2010, where he had been progressed in age. Bin Laden had been attempting to avoid detection from the authorities and some considerable time had passed since any last confirmed face image of him. Therefore, a forensic artist was employed in the FBI to produce an age-progression image of Bin Laden to show what he might look like currently in a bid to try and locate him (see Figure 1.2-2) (The starting image used was from 1998 (BBC, 2010)). The method not only involves the manipulation of the existing features/textures in the image to simulate ageing (Mullins, 2012), but sometimes, in addition, features/face parts are sampled from other face images in order to create a realistic plausible face image, for example with greying hair. However, the forensic artist sampled features/textures from the photograph of a well-known Spanish politician. The facial depiction bore an identifiable resemblance to the politician so that following the release of the age-progression image, the Spanish population noticed the similarity and the comparison was made public (Tremlett, 2011).

Osama Bin Laden (Left), FBI age progression (Middle) and the Spanish Politician, Gaspar Llamazares (right).

It is apparent that the artist directly sourced ‘textures’ from the photograph of the politician so that the photo of Bin Laden could be progressed in age – the hair and forehead have been cut and pasted onto the age-progression image. The beard/stubble also appears to have been sampled in portions along with generic skin textures across the face. It is usual to search for a face image that is similar in textures and contrast to the face being age-progressed so that features/textures are congruent with each other. Even though the samples might be small, it seems that those familiar with the politician (the Spanish public) or specifically with the photograph, were able to detect enough similarities when viewing the age-progression that they were able to identify from whom the features/textures had been sampled. Following on from this, Gaspar Llamazares, the Spanish politician whose image they used, found a second ‘photofit’ type image of another wanted terrorist, the Libyan al-Qaeda ‘number 2’, Atiyah Abd al-Rahman (see

Figure 1.2-3).

Figure 1.2-3: FBI age-progression of Atiyah Abd al-Rahman

Age-progression of Atiyah Abd al-Rahman (left) and photograph of Gaspar Llamarez (right).

Again, it appears that hair and skin textures, as well as the eyes, have been sampled from the face of the Spanish politician and used to progress in age the image of Atiyah Abd al-Rahman in a similar way to the Bin Laden age-progression.

Similarly, within the domain of forensic facial depictions, other types of facial depictions are created, including craniofacial reconstructions (approximating the appearance of a face from the skull) where the process of sampling textures/features from other face images and compositing them onto the reconstructed shape, is also used. Sampling from *one* whole face image is preferable as the lighting and quality remains consistent between sampled portions. However, with this arises the problem of how much and which parts of a face image can be sampled without the depiction looking like the face donor? As can be seen in the FBI age-progression cases, hair, facial hair and skin textures rendered the age progressions identifiable as the known face. In anecdotal reports for craniofacial reconstructions where textures are sampled from face images and added to the shape, the resulting depiction has been reported to resemble the identity from which the textures were sampled (Smith, K. and Shrimpton, S., personal communications, 2017).

Another example of forensic facial depictions is that of the eyewitness composite. This involves an eyewitness describing the face of a suspect/perpetrator in order to create a face image that behaves as a pictorial statement and helps to illicit potential identification leads in an investigation (NPIA, 2009). This can be done with a sketch artist or using one of the automated compositing systems. A few of these systems (Identi-kit, PHOTO-fit, PRO-fit and IQBiometrix' 'FACES' software (Cote, 1998)) use features sampled from real faces for the database from which to choose from (Hasel and Wells, 2007, Frowd, 2015, Davies et al., 2000), the latter two are still used in the U.S.. However, even though these features are placed within a whole face context (composite) it is not known if the facial feature donor's identity is concealed by the process of compositing. More recent sophisticated holistic composite systems such as EvoFIT (Frowd et al., 2011) and EFIT-V (Solomon et al., 2012) use synthesised whole faces that may be less problematic due to not sampling individual features.

1.3 Summary

To summarise, created faces are used in forensic contexts to help with investigations into online criminal activity as well as for depictions that aim to help identify individuals. There is one main question that arises from these two applications of created faces in forensic contexts that is the primary focus of this study: if features/ face parts are sampled from other donor faces and composited to form a new one (created face), can the donors still be recognised within the new identity? For the scenario of fake face profile images, it is likely that the face databases available to the investigating authorities will contain individuals who do not wish to be recognised (e.g. investigating officers). Veridical whole images of the officers from the databases cannot be used in these circumstances as their identity would be compromised as those familiar to them would recognise them. This PhD study investigated if compositing features from face photograph databases still conceals the donors' identity. This question can be applied in a similar way to forensic facial depictions, investigating how features/textures can be sampled from donor faces and composited onto depictions whilst concealing the donors' identity. The study will provide an understanding of how much and which parts of a face can be used in fake face profile images and forensic facial depictions by following a methodology that tests the effect of compositing on face recognition. The compositing technique also serves as a potential way for investigating authorities to generate multiple new face identities for online face avatars to be used in the covert surveillance of online criminal activity, in an efficient way.

Recognition of these types of face composites will be tested in a psychological testing paradigm to try and answer this question and potentially guide practitioners on how to create these composites without compromising a donor's identity. Known (familiar) faces will be used as targets where their features will be replaced with those from other unknown faces, to form stimuli that vary in feature replacement as well as how much of the face has been replaced. The first part of the study will establish which parts of the face may need to be changed by testing for a feature saliency hierarchy. This hierarchy will then inform generation of the stimuli for a second part that will establish how much of a face needs to be replaced to conceal identity through compound replacement of facial features. These stimuli will then be tested in a familiar face recognition paradigm where participants will attempt to recognise and identify each identity, to investigate how these manipulations affect recognition rates.

2 Literature review

In order to be able to investigate whether face donor parts/features are recognisable within a composite, previous research into how humans recognise familiar faces and how psychological testing paradigms can be used to investigate the role of features and face parts within a composite face, will be discussed.

2.1 Context of study

To summarise, composite faces are often created using combined features from images of real faces. These composite faces may be utilised in covert surveillance of online criminal activity (for example, in the context of facial avatars for use on social networking sites), forensic identification (for eye-witness recall or craniofacial depiction), computing gaming (personal avatars) and training purposes (character avatars). The created faces need to appear realistic and may need to pass as an image of a real person. False identities may be created using features from face images for on-line offender surveillance and in these circumstances real faces would also not be appropriate for use. It is known that familiar faces can be recognised from partial images or single features (Jarudi and Sinha, 2003). Research has also shown that the combination of one face with another will alter the recognition of each face (Davies et al., 1977, Young et al., 1987) and the removal of a single feature from a face will reduce recognition (Fraser et al., 1990, Roberts and Bruce, 1988). For example, reduced recognition has been recorded with the removal of eyebrows (Sadr et al., 2003). However, it is not known whether replacing a target feature with an unknown one would render that composite recognisable as the target, or how many elements from other faces are necessary in the composite before the target face becomes unrecognisable. These issues may be paramount to the security of many law enforcement operations, and may also be relevant to the recognition of on-line offenders who may use computer-generated ‘avatars’ derived from an image of their real face. If images of the faces of police officers, military personnel or security agents are utilised for the creation of new face images, the safety of the agents may be compromised if they can be recognised from the composite faces. The results of this study will also help to inform practitioners using compositing for facial depictions in both forensic and archaeological scenarios. The results will provide some guidelines as to how much of a face can be sampled from to provide 'textures' to a facial depiction without the finished imager resembling the face that the textures were sampled from. Therefore, this research aims to establish how much of a target face is recognisable as part of the composite face by studying facial composites produced using familiar target faces.

This literature review will cover the relevant areas of face perception literature that will allow for a comprehensive and robust experimental design as well as the prediction of hypotheses that can be made based on previous empirical research. They are follows:

- Featural manipulations will be used in the first phase of compositing to establish any featural hierarchy and, therefore, previous literature on feature saliency will be discussed
- Holistic processing: The current study will require a familiar face recognition task to occur and therefore the differences between familiar and unfamiliar face recognition will be outlined, alongside the different levels and types of familiarity as well as how to obtain a familiar face stimulus set.
- Whole face processing will be used for recognition and as such, the mechanisms used for whole face processing as well as part-based processing will be explored. This will give an understanding of how recognition might be disrupted depending on how much of a familiar face has been changed through compositing.
- Certain formal aspects of a face image, contrast and spatial frequency, will be discussed in relation to face recognition as these provide cues for recognition that will still be evident in the facial composites.
- Facial characteristics will be summarised to give an understanding of how different types of faces might be recognised compared to each other, in terms of how that particular memory is stored and extracted. This will provide an indication as to whether some face targets may be more susceptible to compositing manipulations than others.
- A brief description of how Automated face recognition systems work will be given and summarised to give an understanding of differences in face processing

compared to humans and so a hypothesis can be made about their performance with the composite stimuli.

2.2 Familiar and unfamiliar face recognition

Empirical research has shown that there are differences in both the quantitative and qualitative processing of familiar and unfamiliar (or newly familiar) faces (Bruce, 1982, Hancock et al., 2000, Barton et al., 2006, Bruce et al., 2010). A familiar face is defined as a face that the viewer has been repeatedly exposed to, most likely over a period of time, where the person has encoded a ‘memory’ for the face in the brain (Johnston and Edmonds, 2009). It is thought that familiarity is on a (potentially non-linear) spectrum, rather than completely binary or distinct, and therefore the processes used in recognition may also fall on a spectrum from that used for unfamiliar faces (more akin to face perception than recognition), to that used for those familiar to us (Campbell, 1999, Campbell et al., 1995, Clutterbuck and Johnston, 2005, Clutterbuck and Johnston, 2002, Collishaw and Hole, 2000, Schwaninger et al., 2002, Veres-Injac and Persike, 2009).

Research has shown that for unfamiliar faces there is a focus of attention to the hair and the external parts of the face, as well as any other identifiable cues, such as facial hair or adornments (Campbell et al., 1995, Ellis et al., 1979, Jarudi and Sinha, 2003, Longmore et al., 2015, Young et al., 1985). The features relied on more in unfamiliar face processing (e.g. hair) are clearly visible from both up-close and longer distances and are therefore easy to focus on to generate an initial face memory. However, those features are likely to change over time and are therefore not robust enough to form a structural familiar face memory (e.g. hairstyles change). Subsequently, for familiar face recognition the focus shifts to look at more invariant parts of the face that are less easily changed and more stable over time such as the internal features (eyes, nose and mouth) in order to recognise someone familiar (Ellis et al., 1979, Young et al., 1985). It is also possible that repeated exposure to a face will inevitably involve communication through movement of the internal features of the face (i.e. speech and expression) forcing our attention to this area rather than the external parts (Ellis et al., 1979). Ellis et al. used familiar faces of famous people to test for the saliency of internal or external features compared to unfamiliar faces and found an internal feature advantage for familiar faces. fMRI studies support the notion that the processing of familiar and unfamiliar faces is different with an external feature advantage for the latter demonstrated through the activation of face specific nodes (see (Natu and O'Toole, 2011) for a review). This is further supported during the encoding process by a study that recorded eye tracking movements during a familiarisation process where participants spent longer focusing on the external features of the face (Henderson et al., 2005) in order to become familiar with the new faces.

Conversely, some studies have not found this pattern of results and found no advantage of external feature encoding and recognition for unfamiliar face perception (Young et al., 1985). Clutterbuck and Johnston (2002) used a discrimination/matching task involving faces of varying levels of familiarity (familiar, moderately familiar and unfamiliar) and internal and external feature conditions. They found external feature matching accuracy was only increased for discriminating different highly familiar faces compared to unfamiliar faces. However, overall they did find an internal feature advantage (faster reaction times) for highly familiar faces compared to moderately familiar and unfamiliar faces and error rates were higher for matching unfamiliar faces compared to familiar ones. In addition, Longmore et al. (2015) used a short familiarisation process with a single photograph and found no advantage for external features over internal features when matching to the same photograph at test. However, an internal feature advantage was found when generalising to a novel viewpoint. Obscuring the hair in a subsequent study further enhanced generalization to novel viewpoints suggesting the external features can detract from attending to the more robust internal features, which would allow for a more stable face representation, during encoding.

Sensitivity to feature displacement (shifting features slightly away from their congruent position) is greater for familiar than unfamiliar faces (O'Donnell and Bruce, 2001, Brooks and Kemp, 2007) with the former researchers finding an effect for the eyes and hair with familiar faces (only hair with unfamiliar) and the latter for the eyes and nose only: this may be the result of a more stable position of the eyes and nose whereas the mouth is dynamically animated so the thresholds for displacement are higher, and the ears may be obscured by hair. It should be noted that when referring to unfamiliar face processing, the research is, in fact, mainly focusing on the discrimination of those faces (Burke et al., 2007, Megreya and Burton, 2006). For example, unfamiliar faces and face parts are often assessed by our ability to discriminate between them (Goffaux, 2012, Logan et al., 2017) and matching tasks may be used to test how similar two unfamiliar faces are (Goffaux, 2012). Encoding and processing studies testing for the role of featural and configural information in relation to an overall holistic processing technique have also often used unfamiliar faces as stimuli (Ramon et al., 2010, Logan et al., 2017, Vesker and Wilson, 2012). These studies are useful in that they tell us something about how we perceive the physical components of a face and in a real life situation we are required to discriminate between unfamiliar or newly familiar faces on a regular basis. However, even with this abundance of unfamiliar face perception research, the results are not always generalisable to a true familiar face situation, which is mostly due to the difficulties in testing familiar faces under laboratory conditions.

Familiarisation

In order to test familiar face recognition, a set of familiar faces is required. This is obtained through the process of familiarisation which refers to the process during which a face is encoded as a memory (Clutterbuck and Johnston, 2007, Clutterbuck and Johnston, 2005, Schwartz and Yovel, 2016, Johnston and Edmonds, 2009, Bonner et al., 2003). Familiarity refers to a) whether the face is known or not, but also b) how familiar the face is, which can be considered to be on a spectrum (Clutterbuck and Johnston, 2002).

There are two methods available for obtaining a familiar face stimulus set in order to test familiar face recognition and recall in an experimental paradigm: naturally familiar faces (faces that have become familiar over a period of time in a natural environment (Longmore et al., 2015)) and Laboratory familiarised faces (a set of faces that are learned within an experimental setting, sometimes over a short period of time and with varying degrees of exposure, format and associated semantic information (O'Donnell and Bruce, 2001, Baker et al., 2017, Schwartz and Yovel, 2016, Wilkinson and Evans, 2009, Bruce et al., 2010)). In some cases, familiar target faces are not available to the experimenter so the process of familiarisation is simulated within experimental settings, to allow a familiar face recognition type task to occur, where participants are shown face images, with varying exposure durations, to promote some immediate face encoding that may result in some kind of memory for that face (Kok et al., 2017, Roark et al., 2006). However, as Hole and Bourne (2010) point out, this encoding session can be very short and unnatural and often no associated semantic information is given. Tong and Nakayama (1999) suggest that a truly robust face representation in the brain requires extensive exposure to the face under various conditions over a period of time and that even after multiple exposure to a new face in order to simulate familiarisation, the participants in their study still performed better when visually searching for their *own* face (naturally familiar). This suggests that even after a fairly extensive simulated familiarisation period where the new face had become 'familiar' to some degree, there were still differences between the processing of these newly familiar faces and naturally familiar faces. Simulated familiarisation techniques may also create an encoding that is very image dependent as the face may have, in some instances, only been familiarised with a few static viewpoints (Roark et al., 2003), which is not in line with a natural way of becoming familiar with someone.

A natural process of familiarisation is not an easy one to simulate in a laboratory setting. There are problematic processes that need to occur that involve trying to familiarise participants with a face set whilst keeping the experiment unbiased towards those faces, generating unbiased associated semantic information and communicating that to participants, familiarisation over various viewpoints and angles and the issue of recall (how will participants indicate that they know who the face belongs to?). Therefore, it can be assumed that in previous studies, unless stated otherwise, participants who have been familiarised with a face set briefly (simulated), are in most cases, not truly familiar with the faces, but rather have a short, non-robust, encoded 'memory' for that face, or in some instances, an image dependent (pictorial) encoding of the face rather than a truly familiar structural memory. There is an argument for using simulated familiarisation in that with a comprehensive familiarisation technique there is greater control over the process and the degree of familiarity is even across stimuli. However, as mentioned before, the process is notoriously difficult and inherently biased, so, where possible the second method of using naturally familiar faces is considered preferable and more ecologically valid.

Still, some disadvantages exist with using naturally familiar faces: differences in the degree of familiarity across a stimulus set (one face may be highly familiar and another face only slightly familiar, even if familiarity thresholds are met), and limitations on targets that can be used to create a homogenous group familiar to a large group of recognisers (participants) that will inevitably result in uneven sample sizes (e.g. some participants may be familiar with twenty faces and others with two). It also seems that varying levels of familiarity not only yields a difference in recognition amplitude between known stimuli but also demonstrates qualitative differences in the processing of faces across varying levels of familiarity: Clutterbuck and Johnston (2002) found an advantage for highly familiar faces in a matching task for internal features compared to moderately familiar and unfamiliar faces (amplitude effect evident in response times) and suggests that the internal feature advantage may serve as an index of familiarity level (Clutterbuck and Johnston, 2007). Additionally, using naturally familiar faces often requires sampling faces from one specific pool (e.g. actors). During the experiment some learning, by the participants, may occur as to which face pool a stimulus belongs to, to enable faster and more successful retrieval of face/identity memory information on the assumption that all faces in the experiment fall within that same face pool (this bias can also be observed when using simulated familiarised faces). Familiar face pools usually consist of a relatively small number of exemplars, reducing the pool size from which participants need to extract a face memory and therefore, perhaps, inflating any recognition rates that are observed compared to those occurring in a real-life scenario.

Another shortcoming when using truly familiar faces, is the lack of control over how familiar the participant is with the target, which may cause noise in the experiment through differences in the amplitude of recognition and possibly some qualitative differences in recognition processing, depending on the level of familiarity. Some researchers have suggested minimum timeframes over which familiarisation was to have taken place to increase the likelihood that participants would be familiar with the targets in the study. Ramon and Van Belle (2016) used personally familiar participants in their study testing whether familiarity enhances global processing of faces using a two-to-one alternative forced choice delayed matching task. Participants were shown both unfamiliar faces as well as familiar ones, which consisted of their classmates. Participants had been studying with their fellow students (n=30) for about two years at the time of testing with some knowing each other for a maximum of five years. Therefore, familiarity, based over a period of time, could be established. As an example, Ramon and Van Belle's research provided a minimum time-frame over which natural familiarisation would have taken place. Face recognition tasks also need to account for the fact that participants may not have ever been familiar with the faces in the experiment. In order to test for this, Frowd et al. (2014) used an experimental paradigm where participants were tested on exemplars (eyewitness composites) of a celebrity and requested for a name in a testing block. This was followed by a block where participants were shown unaltered images (veridical) of the celebrity to test if they were actually familiar with the celebrity in the first place.

Given the literature, it is preferable to use truly familiar faces over a simulated familiarisation methodology due to the quality of natural familiarity as well as for the practical reasons of a) stimulus images are easier to obtain as they do not need to be generated, b) no semantic information needs to be constructed as it already exists in a natural way and c) no extensive familiarisation period is needed that would extend the length of the study and make recruitment more difficult.

Different types of familiarity

Celebrity face images are often used in face recognition tasks as they provide a familiar face set for which recognition and recall type tasks are easily tested and the abundance of celebrity face images on the internet makes for easy access and creation of a (relatively) controlled stimulus set (Carbon, 2008, White, 2004, Lander et al., 1999). However, it is possible that celebrities are a unique type of face set due to their ‘unnatural’ type of familiarisation: they are not learned through personal interaction and there is little surrounding contextual information that would normally be associated with naturally familiar faces such as lecturers. For example, Carbon (2008) demonstrated differences between familiar celebrity images and familiar non-celebrity images (naturally/personally familiar), in particular a qualitative difference where modifying or removing facial hair resulted in a reduction in recognition accuracy for celebrity images but not for the non-celebrities. This small deficit, perhaps, suggests that familiar celebrities fall in between unfamiliar (or newly familiar) and personally familiar faces in terms of the qualitative face encoding and retrieval processes involved. If personally familiar faces are invariant to changes in facial hair, but celebrity faces aren’t, it suggests a more robust memory representation for personally familiar faces where the less stable hair/facial hair feature (hairstyles change) is no longer needed as preference for encoding is now given to the more stable and invariant internal features of the face. It is important to note that Carbon’s study makes a distinction between iconic images of celebrities and other more candid or unseen images of the celebrities, and propose that iconic image familiarity (referred to as iconic processing) is probably based on pictorial encoding rather than a familiarity with the face.

These qualitatively different processing techniques are almost solely due to the familiarisation process involved with celebrity faces (media - images). In contrast, there are similarities between celebrity and non-celebrity familiarisation (different viewpoints, dynamic and static, non-rigid motion (expression), multiple exposures and time-based); the main difference is the lack of one-to-one interaction. This lack of interaction (context) with familiarisation of celebrity faces is fuelled by the sole use of primarily 2D media based (or perceived 3D) for familiarisation, but one could argue that the time/place and emotional state experienced during presentation of the celebrity face as well as the ‘media’ scenario within which the celebrity face is placed, does provide some kind of associated context. If there is a lack of one-to-one interaction with celebrity faces, then a it is possible that performance is better with celebrity faces in a 2D (or perceived 3D) image-based (congruent) laboratory environment compared to trying to recognise a personally familiar face (lecturer) without the necessary contextual information (incongruent) that is so often embedded in the encoding and retrieval of these faces. In support of this, another potential advantage could be that celebrities are often seen in different guises and appearances due to the nature of their work: most celebrities are actors or musicians who require regular reinvention of personal image presentation as well as facial adornments/modification (such as wigs or facial hair) so a it is possible that they would be more invariant to featural manipulations, in particular the hair.

2.3 Feature saliency hierarchy

There has long been considered a general hierarchy of feature saliency that changes according to how familiar the face is. Most researchers agree that the upper part of the face, in particular the eyes, are the most important feature for familiar face recognition (Fraser et al., 1990, Haig, 1986, Schyns et al., 2002, Tanaka and Sengco, 1997): this is based on research that removed the eyes from the face or masked them, or showed them in isolation or even moved their position on a face. In general, the nose has been found to be the least salient. Logan et al. (2017) also found that this general feature saliency hierarchy was constant between embedded and isolated conditions. Eye tracking studies also show a greater deal of saccadic movement to the eyes for static faces with the focus primarily on the centre of the face (Barton et al., 2006, Bindemann et al., 2009). However, the lower face and mouth become more important during conversation and perception of expression (Malcolm et al., 2008). Davies et al. (1977) used compositing (Photo-fit) to replace single features in faces to see how they affect recognition as well as to test for a feature saliency hierarchy. The researchers used unfamiliar discrimination tasks to test the effect of the featural manipulations (eyes, nose, mouth, chin and forehead) resulting in a saliency hierarchy of (from least to most) forehead, eyes, mouth, chin and nose. They found that accuracy was lower for changes made to the lower features compared to the upper ones.

Sadr et al. (2003) found eyebrows to be even more important than eyes for familiar face recognition. They tested recognition of celebrity faces with their eyebrows removed and found bigger decreases in recognition than when the eyes were removed. This study involved targets with fairly distinctive eyebrows (or high levels of contrast with the surrounding face), perhaps biasing the results towards low recognition rates when the eyebrows were removed by disrupting the contrast pattern across the face. White (2004) used a familiar match/mismatch task with pairs of faces to test whether eyebrows masked with a Band-Aid or filled in with surrounding skin would disrupt configural processing more, indicated by response times. He found that response times were slower for eyebrows masked with a Band-Aid in an inverted version (similar to scores found for full faces) compared to the eyebrows erased condition. The inverse of these results was found for the upright version suggesting covering eyebrows with a Band-Aid still maintained configural processing (a small level of contrast is retained due to the difference between the Band-Aid and surrounding skin), whereas if the eyebrows were erased, more feature based processing was adopted as the configuration had been altered. Based on the theory that texture/contrast is potentially more important than configural processing, hypothesised by Burton et al. (2015), one could attribute low recognition rates with the removal of eyebrows to a change in texture/pigmentation or contrast rather than specifically a change in feature shape or removal. Following on from this Gilad et al., (2009) found that the contrast across the face was important for recognition and one could assign this effect to the texture or pigmentation of the face. This theory would support the findings of Sadr et al. that if heavy dark eyebrows are removed it disrupts the texture contrast across the face, therefore, reducing recognition rates. However, as Carbon and Leder (2005) point out, prior knowledge about the reliability of the position and shape of a feature can be made for more stable features such as the eyes (their position and contrast do not change much over time), but the elasticity of the mouth can make it an unreliable feature for processing (Carbon and Leder, 2005b).

Effect of familiarity: For feature saliency, some researchers, for the purpose of testing unfamiliar versus familiar face recognition, have found different effects for the internal versus external parts of the face, thought to have different levels of importance as a function of familiarity (Brooks and Kemp, 2007, Campbell et al., 1995, Ellis et al., 1979, Frowd et al., 2007a). Any feature hierarchy found in literature may depend on the level of familiarity of the participant. For example, Logan et al. (2017) found higher levels of sensitivity to external feature changes for discrimination of unfamiliar faces when features were embedded and altered within a face. This supports the popular belief that there is an external feature advantage for unfamiliar faces. Following on, sensitivity was higher for changes to isolated features, compared to the embedded feature version, suggesting that sensitivity to featural changes is impeded by what is thought to be dominant holistic processing that does not allow us to attend to specific featural changes in the face. However, during this study they did find a significant feature advantage for the nose in the embedded version; sensitivity was found to be higher to changes made to the nose compared to other internal features, with the eyebrows showing the lowest sensitivity. This suggests that the effect of holistic processing may not be even across features. Eye sensitivity was found to be intermediate and may have been the result of using unfamiliar participants, where the internal features are less important. It should be noted that the resulting feature order, in terms of discriminative sensitivity, was the same for both embedded and isolated versions, although to different amplitudes, suggesting no qualitative differences in processing.

In an unfamiliar face discrimination study by Vesker and Wilson (2012), a condition where only the eyes and one other feature remained in the schematic face stimuli yielded no significant difference between the feature conditions (nose, mouth and head outline) for sensitivity to inter-ocular changes. However, as shown in the previous literature above, eyes have been considered the most salient, so it is possible that keeping the eyes throughout all the conditions provided configural cues and more information in order to discriminate evenly across other feature conditions. This was an unfamiliar discrimination task, requiring participants to observe differences between unknown faces rather than recognition of them. However, the three feature conditions did yield higher sensitivity scores than for the isolated eye condition, suggesting that those features do enable clearer observations of differences between faces (in this case inter-ocular distance), in other words a face context effect: in particular the facilitation of configural processing. The authors suggest that the nose and head outline may have been providing inter-ocular cues because they lie on the same horizontal line as the eyes (relative changes to inter-ocular distance).

A 2016 paper by Abudarham and Yovel (2016) used feature compositing, but using unfamiliar faces for face matching/discrimination tasks. Their research focused on changing faces (adjusting features) to move them around the theoretical model of face space (Valentine, 2001) along its various dimensions that are associated with physical elements of the face, such as features, (e.g. Lip thickness (see Face memory and storage for more detail)). The face space model theorises that faces can be adjusted within their own subspace without a change in identity, but once a face is adjusted more than within the subspace threshold it moves out of its own subspace and forms a new identity. Participants were required to make judgements about these adjusted faces, such as, 'which face has the thicker lips?' A series of experiments aimed to assess which features participants were most perceptually sensitive to changes of, and from that, hypothesised that these would be the features critical for identity. Their results over various experiments showed high perceptual sensitivity (PS) to lip thickness, hair colour, eye colour, eye shape and eyebrow thickness. Low PS features were found to be mouth size, eye distance, face proportion, skin-colour and nose size. A subsequent same/different task showing the veridical face and adjusted face suggested that changing high PS features resulted in a change in identity, but changing low PS features did not, within their experimental paradigm. They also observed that smaller changes to high PS features resulted in more of a change in identity compared to larger changes to low PS features. These low PS changes were perceptually detectable but were ignored by participants during the same/different identity task. The authors suggest that the high PS features are relied on more for face identity due to their more invariant nature to elements such as viewpoint and expression, therefore changing their shape is more disruptive to identity matching. Hill et al. (2011) suggest that some changes can be made to a face without having any effect on identity whereas other changes will, demonstrating how manipulations can render a face more similar or dissimilar to its veridical self. Their methodology defined unknown faces in a physical face space and used Principal Component's analysis to metrically change faces along different dimensions to find criterion points from old to new identity using a same/different task. The choice of principal component dimensions to be manipulated was randomised throughout the experiments. Their results were based on more holistic changes to a face in order to establish a criterion point, rather than specific featural manipulations.

From the literature, there seems to be varying reports on a feature saliency hierarchy but most agree that the eyes are considered the most salient for familiar faces. It is also worth pointing out the degree of information available from each feature. For example:

- The eyes contain both contrast information (dark pupil, white sclera etc.) as well as shape information. They are dynamic (expression) but the canthi positions remain constant during expression and throughout life.
- The eyebrows, conversely, mainly consist of textural/contrast information. Highly dynamic and can be easily cosmetically modified.
- The nose could be considered mainly an area of shape information with a small amount of contrast information available from the nostrils and any shadows created by the overhang of the nasal tip.
- The mouth, like the eyes, contains both contrast and shape information. Highly dynamic.
- The contour or facial outline is mainly an area containing shape information.
- The hair, an area made up mostly of texture, provides contrast information with some volume shape occasionally available.

The role of each feature within communication may also increase its saliency. For example, the facial contour is largely inexpressive and any communication is as a result of distortions created by the mouth. The nose, again, a stable feature of the face, provides very little communicative information. In addition to the eyes being considered the most salient for recognition, the eyes, as well as the eyebrows and mouth, are the most expressive and dynamic features of the face and considered most salient for expression (Wells et al., 2016).

2.4 Holistic processing

It is thought that faces can be described in terms of their “parts” (features) and that these combined with the arrangement of features, together, make up a whole face. Francis Galton made this very observation back in the 19th century in his book, ‘Inquiries into the human faculty and its development’ (Galton, 1883) by writing, *“…a face is the sum of a multitude of small details, which are viewed in such rapid succession that we* *seem to perceive them all at a single glance",* and summarized this with "..*the whole is truly greater than the sum of its parts*". This section will explore the three main contributors of holistic processing: 1. the parts and wholes effect where a whole face image induces holistic processing, and 2. the role of features and 3. configuration that contribute to holistic processing. These will be discussed in sections on Parts and Wholes followed by Featural and Configural processing.

Parts and Wholes

Face recognition research has looked closely at the role of face parts and how they behave within a whole face. In Tanaka and Farah's study (Tanaka and Farah, 1993) participants learnt whole faces and were subsequently tested in a two-alternative forced choice recognition task with three different conditions: 1. the target feature in isolation (plus foil), 2. the target face again (the foil is the target face with the foil feature) and 3. The target face with the distances between the features altered (foil face is the same as for 2.). Results showed that participants were better at correctly choosing the target face when shown whole faces than when shown target features in isolation. Tanaka and Simonyi (2016) point out in their review of the part/whole paradigm that "*the results indicate that our memory for a single part of a face is embedded in our memory for the entire face*". Non-face stimuli, such as houses, *do* show this same pattern of holistic processing and an advantage for whole object recognition, but to a far lesser degree than for faces (Tanaka and Gauthier, 1997). However, some have argued that this advantage for whole faces may in fact be due to an encoding specificity where whole faces are learned and recognition rates for isolated features are lower simply because they do not match our whole face memory whereas whole face images provide a better retrieval cue (Gauthier et al., 2009). However, there are additional effects of inversion and scrambling that yield equal recognition rates for both whole and isolated parts, supporting the role of holistic processing that only occurs for upright whole faces. Conversely, Leder and Carbon (2005) reversed the parts/whole task by familiarizing participants with isolated feature parts and then testing them on their recognition of these features when presented in whole faces. During test, recognition rates for isolated parts (congruent to the learned format) were considerably higher than when the isolated features were tested in their whole face (incongruent to the learned format). This suggests some kind of whole face interference and dominant role of holistic processing. fMRI studies have also found activations for the occipital face area and fusiform face area during presentation of isolated features (Henriksson et al., 2015). It should also be noted that prior knowledge of the general arrangement of a face (a top-down effect), allows us to make assumptions about parts of the face that may be missing from view (Troje and Bulthoff, 1998) (e.g. part of one side of the face is missing in a three quarter view so an assumption is made about the missing side based on prior knowledge that faces are, in general, mostly symmetrical).

The compositing of features from known faces has been investigated in a previous study by Cabeza and Kato (2000) when face parts from more than one face were combined together to form a novel face that resulted in poor recognition rates for those constituent parts suggesting that the composite was depicting a whole new identity. Participants briefly learned new faces and were subsequently tested on featural composites that contained one feature from one of the learned faces and tested in an old/new paradigm. Cabeza and Kato's study may in fact be a test of localised image discrimination and comparison and the old/new paradigm is a familiar 'recognition' task rather than one of recall (of a specific identity and extension of recognition). In a study by Ramon et al. (2010), the primary focus of which was to test for holistic processing in an individual with acquired prosopagnosia, non-impaired participants were simultaneously tested as a control. Using unfamiliar composite face stimuli, a two-alternative forced-choice paradigm was used to test if swapping features from a target image with those from another face (nose, mouth and eyes) would affect recognition. Participants were probed with a short exposure duration to a target face and subsequently presented with two faces: the target face and a composite (distractor) face, differing only in one feature at a time. A further condition of isolated features was also tested, generalising from the full face probe, as above, followed by two isolated features: one distractor and one target. As expected, control participants were better (correct responses) at the whole face condition during test (congruent to the learned probe) with additional shorter response times, supporting the whole face holistic processing hypothesis. An advantage was found for the eyes in the embedded whole face condition, suggesting that not all features are affected equally by the holistic processing effect, and that this advantage is only found when the eyes are part of a whole face (holistic) rather than in isolation.

In another unfamiliar discrimination task, again this did not test face recognition but rather differences between faces (in this case inter-ocular distance), Vesker and Wilson (2012) found that sensitivity to changes in synthetic full faces is higher than for faces where only the eyes and one other feature are included in the image (not a complete face). In part, this provides some insight into the summative role of all features being included in a face, regardless of their ability to provide face detail in isolation with results showing an amplitude effect for the full-face context.

Composite Face effect

Face composites are often used as stimuli in face recognition tasks, one such design developed by Young et.al (1987) combines the top half of one familiar face with the bottom half of another, referred to as the Composite Face Effect (CFE). Sometimes these halves were aligned and other times they were offset slightly by a short distance. Reaction times for identifying the two individuals were considerably higher for aligned composites than for non-aligned composites. As pointed out by Tanaka and Simonyi (2016) in their review of facial parts and wholes, this may be due to our inability to attend to specific parts of the face without ignoring the surrounding face. Several other researchers have used the composite face task to support the holistic face processing hypothesis, with variations on the methodology outlined above and for both unfamiliar and familiar faces (Young et al., 1987, Le Grand et al., 2001, Michel et al., 2006, Jiang et al., 2011, Laguesse and Rossion, 2013).

Ramon et al. (2010) repeated the composite face task and unlike most other experiments that only change one half, they changed *both* the top and bottom halves of the composite as an extra condition to test whether an effect was caused by alignment or by a change in identity. A same/different task was used for the upper halves of the face, but with the added difficulty of a different bottom half in half the trials across misaligned and aligned conditions. They found no effect of alignment when the bottom halves of the faces were the same but a large effect of alignment when they differed suggesting that a whole new identity was being formed when the bottom half was also different. Using fMRI, Harris and Aguirre (2010) used a composite face paradigm (top and bottom halves of two different faces composited and aligned) to test for holistic and feature/part based processing and found activations for both processing mechanisms within the fusiform face area (misaligned composite halves still activated the fusiform face area suggesting face part based processing was also taking place here) and Freiwald at al. (2009) also found evidence of part based processing using Macaques.

The composite face task, therefore, demonstrates our inability to attend to just one part of a face within a whole face context reinforcing the argument for holistic processing. The whole/parts task effect is also found in both unfamiliar and familiar types of face recognition (Tanaka and Simonyi, 2016). Young (1987) initially demonstrated this effect with familiar faces but it has also been found with unfamiliar faces using a face-matching task (Hole, 1994) suggesting that this may also be an encoding issue. Inversion, however, reduces the power of the composite face effect. The lack of configural processing with an inverted face allows us to attend to specific parts of the face, i.e. the two different halves (Young et al., 1987). To summarise, faces appear to be processed holistically when a whole, aligned face is presented that prevents us from attending to specific parts of a face. It seems that if a face image contains the whole configural pattern of features in alignment, then holistic processing is activated.

Featural and configural information

The composite face effect outlined above, demonstrated that misaligned face halves were more easily recognised as their respective identities compared to when they were aligned, as the alignment of the configuration of the features created a whole new face identity that prevented attending to specific parts of the face. In addition, research has also shown that the *specific* configuration, or distance between features (e.g. 1cm between the nose and mouth), also contributes towards the holistic processing of faces. Previous research had suggested that the configuration of features, or more specifically the *global* arrangement of features (first-order relations) and their respective distance between one another (second-order relations), is incredibly important for familiar face recognition (Collishaw and Hole, 2000, Sergent, 1984, Young et al., 1987, Tanaka and Sengco, 1997). It is thought that familiar face recognition relies on a combination of these second order relations as well as *local* feature shape (featural) information to generate an overall image of the face, often referred generating a ‘gestalt’ (Maurer et al., 2002, Mondloch et al., 2010, Mondloch and Maurer, 2008). Bartlett and Abdi (2006) point out that when researchers refer to configural processing, they are most likely referring to the second order relations, or rather spacing between features and that one change in distance, does in fact change the spacing of the rest of the face. Baenninger (1994) makes the distinction that locational configuration (the correct 1st order relations of a face; two eyes above a nose, above a mouth etc.) details are more disruptive to recognition when altered than relational configuration (2nd order relations of a face). Goffaux (2012) goes on to use the term “IFP” when describing configural changes as Interactive Feature Processing. Although these changes may be small, research has shown that we are sensitive to even very small changes in distances between features (Haig, 1984, Barton et al., 2001, Brooks and Kemp, 2007). Tanaka and Sengco (1997) found that adjusting the spacing of, for example the eyes, between learn and test in newly familiar (or learnt) faces, resulted in lower recognition rates, again supporting the importance of configuration in face encoding and retrieval.

In addition to configural information, as mentioned above, research has shown that *local* featural (feature shape), also contributes to holistic processing. The relative importance of configural (global) and featural (local) information has been contested throughout the literature. Ramon and Van Belle (2016) used a two-to-one alternative forced choice delayed matching task where stimuli were degraded (blurred) and the similarity/dissimilarity of the face pairs was altered to test for adopted processing techniques and whether familiarity facilitated either global (configural) or piecemeal (featural local) processing. They found that familiarity was associated with an experience-based enhanced global processing technique, with less reliance on local or feature-based processing. However, Schwaninger et al. (2002) repeated Tanaka and Sengco’s type of study for both familiar and unfamiliar faces (or newly learnt faces) and found that there was no qualitative difference between them and that they both relied on featural and configural information depending on the type of image being shown (scrambled or whole). Another study (Sandford and Burton, 2014) also found no difference in accuracy between unfamiliar and familiar faces when testing the role of configuration by asking participants to correct spacing distortions that had been applied to face images.

So far, many researchers have presented evidence for a dual-hypothesis route where both configural and featural information is needed for face encoding and retrieval and perhaps a slightly more important role for configural information (Cabeza and Kato, 2000, Tanaka and Sengco, 1997, Favelle et al., 2011, Frowd et al., 2014, Gilad-Gutnick et al., 2012). Collishaw and Hole (2000) provided a comprehensive review of this by using disrupted images to isolate the two processing techniques: accuracy was still above chance when only one route to recognition was available and when both routes were disrupted recognition failed, supporting the dual-hypothesis route. Baenninger (1994) used an unfamiliar target present/absent task to test for the separate roles of featural and configural information that is used by children in a comparative study with adults. Target faces either had their features configurally altered, or the features were removed (with one feature remaining). Results suggested little qualitative differences between adults and children and that both relied more heavily on configural processing. Le Grand et al. (2001) isolated the two types of information by altering face images through moving the position of features (configural) and swapping the features (featural-shape) and testing identity thresholds using a same/different task for unfamiliar faces (see Figure 2.4-1). They found no differences in accuracy for both stimulus types when using ‘normal’ participants, however, they found, as expected, a deficit for the configurally altered stimuli when they were presented inverted.

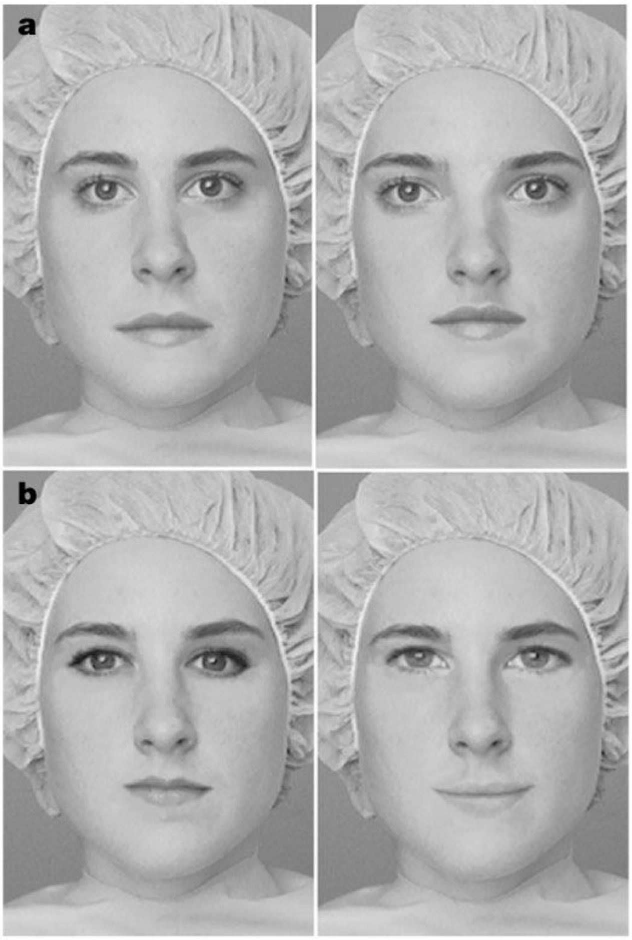


Figure 2.4-1: Le Grand et al. study using altered composites

Example of stimuli from the study that tested configurally altered composites (top row) and featurally altered composites (bottom row) using unfamiliar faces. No differences were found in accuracy using a same/different task except for when faces were inverted. (Le Grand et al., 2001)

Yovel and Kanwisher (2004) also used a similar paradigm of manipulating either configural or featural information but creating several manipulated exemplars of unfamiliar faces for a discrimination task looking at activation of the fusiform face area using fMRI. For example, one face would generate four stimuli where the configuration was the same, but the features differed and four stimuli where the features remained but the configuration was altered in a comparative study against non-face objects to test for domain specificity and found fusiform face area activation for faces only but no preference for configural or featural information. Freire et al.'s study (2000) aimed to use a similar paradigm of isolating configural and featural changes but to investigate each processing role in encoding and retrieval by using a delayed forced-choice matching task for the latter: the short duration of encoding (a few seconds) rendered the images unfamiliar/newly familiar. Furthermore, fMRI studies of prosopagnosic individuals (inability to process faces configurally) have found a larger activation for the left hemisphere and featural processing preferences when the right hemisphere has been damaged suggesting that featural and configural processing are processed on separate sides of the brain (Yin, 1970, Marotta et al., 2001). See (McKone and Yovel, 2009) for a review of research exploring the role of configural and featural information.

Recently, however, Burton et al. (2015) suggested that second order relations that are used for configural processing were perhaps not as important as once thought. As pointed out by the researchers, previous findings that support the importance of configural processing, have suggested that an inverted face reduces recognition rates considerably and hypothesised that the reason for this was because the first and second order relations of the face were disrupted and thus an overall gestalt of the face could not be constructed. However, significant disruptions to the second order relations are made when a face image is stretched considerably, in a linear fashion, on its vertical axis (Hole et al., 2002) without greatly reducing recognition rates demonstrating a lesser important role for configural processing. Stretching (vertically stretched: appearing to view the face image from a more ¾ view) has been found to even improve recognition rates of eyewitness composites that may help to remove any inaccuracies from the compositing process (Frowd et al., 2014, Davis et al., 2015, Frowd et al., 2013). Frowd et al. (2014) used perceptual stretching not only for recognition of the composites, but also for the veridical images of the celebrities from which the composite was made. They found an effect with a positive increase in correct naming for physically stretched composites in comparison to the non-stretched condition but no advantage for stretched celebrity target images in comparison to their veridical counterpart. In contrast, other types of transformations do reduce recognition rates and the same researchers found that slanting an image to the left or right slightly impaired our ability to recognise a face (Hole et al., 2002).

To summarise, researchers have contested the relative importance of featural (local) and configural (global) contributions towards the holistic processing and identification of faces. In general, it appears that most studies found both streams of information contribute in some way. However, Burton et al. (2015) continued on to suggest that perhaps what affects recognition rates the most is actually an issue of pigmentation/texture or colour and that the contrast patterns of these (dark pupils against a white sclera) are important.

Facial contrast and pigmentation

In addition to, and in some ways a contributor to, the featural and configural streams of information available from the face, is facial contrast. It is thought that face recognition is activated by contrast patterns (otherwise known as spatial frequency) across the face, i.e. the darkness of the eyes, nostrils and lips, with the surrounding lighter skin and the contrast of the iris/pupil against the sclera (Keil, 2009, Costen et al., 1994, McNeill, 1999). Neurologically certain neurons are activated by contrast, that may aid the recognition process by summarising the contrast/spatial frequencies of the face (Keil, 2009). Spatial frequencies are often referred to as how often the light changes as it passes across the face width and can be described as cycles per face width (Hole and Bourne, 2010). Sergent (1986) found that certain frequencies of information might be useful for different types of encoding. For example, high frequency information such as the edges of features and detail may provide cues as to the featural properties of the face, whereas low frequency information such as contour and overall shape and shadow may provide more configural information.

Dakin and Watt (2009) demonstrated that the horizontal structure of the face contained more useful information than the vertical by selectively removing horizontal and vertical spatial frequency information using the analogy of facial contrast as some kind of ‘facial barcode’. Nasanen (1999) found that using a simple unfamiliar recognition task with a short familiarisation phase and face images altered with a band-pass filter to adjust the spatial frequency, the optimum image for correct recognition was 10 cycles per face width with a Gaussian filter and bandwidth of 2.0 octaves. High frequency face images, such as line drawings, can still be recognised in face matching tasks, however, they falter when encoded prior to a subsequent recognition task (Leder, 1999). The task does becomes easier when some low frequency information, such as shading, is available (Bruce et al., 1992). This strengthens the argument for our need for shape from shading information (perceived 3D shape) in order to recognise a face (Kemp et al., 1996). It is thought that the lack of positive 3D shape (naturally congruent) information in negated images (i.e. reversed) is responsible for our difficulty in recognising faces under these conditions (Russell et al., 2006, Bruce and Langton, 1994, Bruce and Young, 1998). The researchers hypothesise that normally a face is lit from above as in a naturalistic setting where most light sources are above so when a face image is negated all of the contrast information is reversed. Johnston et al. (1992) played around with reversing the lighting for both positive and negated faces and found that negated faces lit from below (the reverse of natural lighting) where more easily recognised than negated faces under natural lighting. This effect may occur due to the disruption of shape from shading information available to the viewer (Hill and Bruce, 1996) arguing the case for our system's need to extract perceived 3D information from faces and face images and Burton et al. (1999) have shown that negation highlights our need for both line and shading information. Hole et al. (1999) used negation and a chimeric composite face paradigm to test for both negation and its role within face processing and found an effect for both the composite face illusion as well as for inversion suggesting negation does involve some kind of configural or inter-relational processing.

Other surface information available from the face is pigmentation (or reflective colour) and it is thought that colour may not be that important for face recognition. Studies have shown greyscale to be just as effective. According to Kemp et al. (1996) and Bruce and Young (1998), removing colour pigmentation does not drastically alter recognition rates of known faces. The researchers argue that colour does not affect shape-from-shading information processing because it does not require it.

2.5 Face Types

This section will outline how individual faces exhibit facial characteristics and could be classified into different face types. Literature on why these different face types are more or less likely to be recognised will be discussed.

Face memory and storage

There are two main theories of how faces are stored in memory: exemplar and prototype based models, which form part of Valentine's theoretical Multidimensional Face Space model (MDFS) (Valentine, 2001, Valentine, 1991). Put simply, Valentine theorises that each face is made up of a number of values and those values each have their own dimension. For example, one value may represent lip thickness which would result in a lip thickness dimension where faces with thin lips would lie on one end of that dimension and faces with a thick lips on the opposite end. The two versions of the model are as follows:

1. The Exemplar-based model theorises that we store each face as its own representation within a 'face space'. More similar faces are clustered together and more distinctive faces further apart but with no central norm to the space.
2. The Norm-based model suggests that we have a multi-dimensional face space where an average or prototype face is in the centre and all other faces radiate out from that. The theory is that more average faces look more similar to one another and are thus clustered towards the centre of the face space around the absolute average. More distinctive faces are situated further away from the average and are less densely clustered.

Lewis (2004) went on to estimate how many dimensions, within face space, are needed to be able to sufficiently describe individual faces from the same race, suggesting 15-22 dimensions would be required.

Research supports the notion that distinctive faces are recognised better than more average faces (Bruce, 1998, Hill et al., 2011, Lee et al., 2000, Valentine, 1991, Valentine and Bruce, 1986, Valentine and Endo, 1992). Additionally, during face/non-face classification tasks, it takes longer to classify a distinctive face as such, and less time for an average face (Valentine and Bruce, 1986). This phenomenon can be explained by Valentine's theoretical MDFS Model (Valentine, 2001, Valentine, 1991) where average faces are clustered together and are therefore easily classified as a face, whereas distinctive faces are further apart from one another and less likely to have many faces around them to support the trigger of a 'face' classification. Logan et al. (2017) conducted a study to assess if the distinctiveness level of a face affected recognition rates for both embedded and isolated features. Using a discrimination task based on different distinctiveness levels for four identities results showed no qualitative differences in processing for featural manipulations: the feature hierarchy remained constant across faces for both isolated and embedded (whole face context) conditions across all identities and distinctiveness levels. This suggests that feature saliency is not affected by overall distinctiveness levels. However, this was an unfamiliar discrimination task and the results may not be generalised to a familiar face recognition. One type of image manipulation that provides an example of the exaggeration of facial distinctiveness is, caricaturing. Caricaturing has been shown to improve recognition rates where veridical/composite face images show lower recognition rates than their counterpart caricatures (Robert, 1999, Frowd et al., 2007b, Lee et al., 2000, Rhodes, 1996, Benson and Perrett, 1991, Benson and Perrett, 1994). It is thought that the process of caricaturing enhances distinctive elements of the face, thus making it easier to extract the memory for that person.

Similarity and dissimilarity of faces

Valentine's model (see section 2.5 Face memory and storage) shows that similar faces are clustered together and dissimilar faces further apart on the various dimensions of the face space (Valentine, 1991). Therefore, any face recognition task can be made easier or more difficult depending on the difference between faces being tested. This becomes especially important in studies involving compositing faces together. For example, it would be unwise to ask participants to discriminate between a face of a 70yr old and a 20yr old in a basic face matching task as the answer is obvious and age is being used as a cue to discriminate rather than the face itself. Previous researchers have matched faces based on visually derived semantic codes, as well as other facial characteristic cues, and contrast and luminance. White (2004) matched pairs of faces for a study on the role of eyebrows by using similar ages, hair colour, and ethnicity. In previous research Ramon and Van Belle (2016) also matched for eye colour and overall luminosity and Goffaux (2012) also matched pairs for luminance. Following on from this, it is important to discuss what changes need to be made to a face in order for it to present a different identity (or rather conceal its original one)? Hill et al. (Hill et al., 2011) address this very issue, suggesting that some changes can be made without having any effect on identity whereas other changes will, demonstrating how manipulations can render a face more similar or dissimilar to its veridical self. Their methodology used Principal Component's analysis to metrically change faces along different dimensions to find criterion points from old to new identity.

Other-face effects

Valentine (1991) suggested that his MDFS model (see section 2.5 Face memory and storage) illustrated a personal face space specific to that person based on their experience of faces they had seen over a life-time, therefore the space was optimal for distinguishing between faces similar to those they had seen: in one's lifetime they may be exposed to a majority of same-race faces. The other-race effect (or cross-race effect) can be defined as the difficulty in recognising faces that are not of the same race (perceived ethnicity) to one's own (Meissner and Brigham, 2001, Mondloch et al., 2010). Further research went on to support a new hypothesis based on the norm-based model indicating that the norm or 'prototype' face may be specific to the race of the owner of the face space (Valentine and Endo, 1992) making any 'other' race faces difficult to process and differentiate due to the vectors heading in a similar direction. The author goes on to explain that the exemplar based version of this suggests other-race faces are difficult to recognize due to being located far away from typical own-race faces in a pseudo distinctive cluster. Research has found that there are qualitative differences in the processing of own and other race faces (Pezdek et al., 2012, Meissner and Brigham, 2001, Brigham and Malpass, 1985, Michel et al., 2006), with a weaker effect for highly familiar faces due to a more robust face memory. In Pezdek et al.'s (2012) study, participants viewed faces either individually, or in arrays of three, with either two race congruent or incongruent distractors. Participants performed worst at a recognition task if other-race faces were presented in an array, rather than individually, and if the distractors were race-congruent, demonstrating that viewing conditions, i.e. real-world environment, affect the other-race effect. Mondloch et al. (2010) used a similar methodology where single features were replaced in a face to test for the effect of featural manipulations using a same/different task and the premise was to test for this effect on both own and other-race faces. They found an advantage for own-race faces when features were swapped. It has been proposed that also there exists an own-gender bias for face recognisers where female recognisers are more accurate at recognising female faces (Loven et al., 2011, Herlitz et al., 2013, de Frias et al., 2006). This is thought to occur due to longer attention to female faces at the encoding stage that results in higher recognition scores for female face stimuli (Loven et al., 2011), but this may also be influenced by the viewers own face space. Through a meta-analysis of literature on sex differences in face recognition, Herlitz and Loven (2013) also found that females remember more faces in general, compared to males. An eye-tracking study found longer fixation on the eyes for own-gender faces in female recognisers (Man and Hills, 2016). Additionally, there is some evidence of an own-age bias with participants better at recognising faces of a similar age to their own (Wiese et al., 2013, Rhodes and Anastasi, 2012).

Facial characteristics

Valentine's model of face space has shown that the distinctiveness level of a face affects our memory and recognition of it (see section 2.5 Face memory and storage): The more distinctive a face is, the more memorable it is and therefore it is more likely to be recognized as it can be easily extracted from a less densely populated area of the face space. Some other characteristics affect our memory for faces. Whether a face is perceived as attractive or not has been linked to how average or distinctive the face is within the face space. In short, more average faces, or faces that sit closer to the population norm, are considered more attractive than faces that sit further away from the norm and are considered more distinctive ((Hole and Bourne, 2010, Rhodes, 2006, Thornhill and Gangestad, 1999, Langlois and Roggman, 1990). Therefore, it is possible that attractive faces are less likely to be recognized than distinctive ones, especially if alterations are made to any face images. Following on from this, attractiveness levels have been linked to levels of trustworthiness. A study by Kleisner et al. (2013) asked participants to rate faces with different eye colours for trustworthiness. They found that faces that were considered trustworthy (e.g. the types of faces associated with brown eyes) were also found to be rated as more attractive. Sofer et al. (2014) conducted a study where participants were asked to rate faces on a Likert scale for trustworthiness. They found that the more typical (average) the face was, the higher the trustworthy rating, demonstrating a link between trustworthiness and distinctiveness: more distinctive faces were considered less trustworthy.

Research initially demonstrated that identity and perceived 'gender' (or sex appearance) information were stored and processed separately as a visually derived semantic code (Bruce and Young, 1986, Ellis et al., 1990) where 'perceived gender' provided no cues as to the identity of a face. However, it has more recently been suggested that the sex appearance of a face can provide some cues as to the identity of the face (Goshen-Gottstein and Ganel, 2000, Ganel and Goshen-Gottstein, 2002) based on the representation for that person stored in the brain through neuropsychological case studies of individuals showing only a single dissociation: a patient could no longer recognise faces but could still perceive gender, the reverse of this was not found and therefore lacking the evidence of a double dissociation that would normally support parallel processing. Additionally, their 2002 study was conducted supporting the single route hypothesis by demonstrating that providing gender specific names to an androgynous target face resulted in longer decision making times for rejecting gender congruent distractor faces suggesting that gender was, in this case, providing a cue for identity: see (Hole and Bourne, 2010) for a review.

2.6 Automated Face recognition systems

In addition to human face recognition in society, developments in technology have allowed for computer systems to replicate this process so that the recognition of faces can be automated. These systems sometimes mimic the face processing mechanisms that humans use by using image processing and pattern recognition techniques in order to match and identify faces (Olszewska, 2016).

Most automated face recognition systems (AFR) use a three-stage approach: first, the system needs to detect that there is a face in the image and to locate that face. The next step is to extract the face or facial features from the face portion of the image and lastly the face/features are classified and matched to a database of faces for recognition (Olszewska, 2016, Chellappa et al., 2010). The last stage may take the form of three different tasks: 1. Verification, where the face image is checked to see if it matches a known target face, 2. Identification, where a person's identity is determined from the face image and 3. A watch list, where the recognition system establishes if the face image is from a watch list and identifies that person. The training sets that the systems use for this process vary considerably in the range of images collected for any one face with respect to pose, expression, age, illumination etc. (Parmar and Mehta, 2013). Some of the different processing strategies for extracting face information from the training images as well, as the novel input image, include:

1. The holistic method where the whole face image is extracted and normalised to a template and then the closest matching face is identified from the database (an example of normalised faces is the Eigenface system (Turk and Pentland, 1991)).
2. The extraction of facial features which are then fed into a classifier system. Feature extraction may use edges, lines, curves, or using a feature based template. For the facial feature extraction stage, the shape and size of the facial features are located and extracted in various ways and sometimes additional configural information is also collected. Chellappa et al. (2010) go on to describe how features can be extracted by marking off key landmarks (points) and/or using shape or texture, or even treating each feature as an individual component. A matching threshold is used to classify an input as either a match or not, with respect to the combined score of each feature classification (match).

3. The hybrid method, that uses a combination of the holistic and feature based methods above, but is more often used for 3D face data. Therefore, automated face recognition systems vary in their processing strategy for extracting and matching faces and facial features, as well as the training set that they use for verification/identification, and this has resulted in varying levels of success. For example, the feature extraction based method is highly affected by changes in pose, as the features are no longer of the same shape and structure as that from the learnt image. Therefore, these types of systems work well for congruent learned and test images (usually frontal) but performance drops for other viewpoints and poses (Parmar and Mehta, 2013).

Humans are remarkably good at recognising familiar faces despite a vast array of different conditions under which faces are seen such as pose, illumination, expression, age and viewing distance (Sinha et al., 2006). It seems that automated face recognition systems are perhaps not as good at this task, with some conditions making for difficult circumstances under which the system is meant to recognise a face (Ring, 2016). One particular group of researchers from the University of Texas at Dallas have extensively tested human recognition performance against research and commercially based automated face recognition systems data that has been collated by the National Institute of Standards and Technology (NIST) in North America. In their 2014 review (Phillips and O'Toole, 2014) of research that has compared humans and computers since 2005, they summarised that for matching of frontal still face images to systems trained on a set of faces (matching for verification), computer algorithms perform better than humans, but the opposite is true for other angles and dynamic footage. Similarly, O’Toole and Phillips (2015) report that recognition of frontal images is the same as humans for unfamiliar faces tested against a database of faces. From the research, automated systems, in general, perform worse than humans for face images that present different or more extreme levels of pose, illumination, expression, poor image resolution and vast differences in age (Chellappa et al., 2010, Phillips and O'Toole, 2014).

As pointed out by O’Toole and Phillips (2015), the underlying algorithms used in these automated systems are mostly unavailable for commercial systems due to their proprietary nature. Therefore, it is not possible to analyse the algorithms in order to understand exactly how they are processing face images as part of the recognition process, unless disclosed by the authors in some cases. However different tasks have been developed over the years in order to test these systems, and from the results some inferences could be made as to how the systems may be processing the images. Some systems, like the Twinsornot website (Twinsornot), compute similarity ratings between two images. For example, two images can be uploaded to the website and a similarity rating, expressed as a percentage, is given. The Face recognition comparison group study by O'Toole et al. (Phillips and O'Toole, 2014) used a comparison of image test, where two images were compared and similarity ratings given, between automated face recognition systems and humans to compare performance. For the easy pairs condition (obviously different or same people), all systems outperformed humans. For the difficult pairs, only half of the systems outperformed humans. In addition they used another standardised test called, 'the good, the bad, and the ugly face recognition challenge' (Phillips et al., 2011) where images of identities under different image conditions, such as illumination and pose, are tested (this translates as testing good images, bad images and very difficult images). Humans outperformed the automated systems for good images, but were only comparable for the bad and very difficult images.

Some researchers have shown that computerised systems show some success when identifying faces where parts have been obscured, otherwise known as faces in disguise. One particular study yielded an identification rate of 55% when faces were obscured with a hat, scarf and glasses (Singh et al., 2017). Upon removal of the glasses this rose to 69%, suggesting an importance of the eye region in the systems processing strategy. The hat and scarf obscure the external parts of the face, whereas the glasses obscure an internal feature, which suggests that feature extraction is being used. Dhamecha et al. (2014) carried out a study where their computerised face recognition system was tested against humans for disguised faces. Their stimulus set contained faces with naturally occurring disguises such as facial hair, different hairstyles, facial masks (e.g. doctor's mask) and hair coverings such as Hijabs. Unsurprisingly, they also found that human participants' performance was better if they were familiar with the face, in line with face perception literature (see section 2.2 Familiar and unfamiliar face recognition for more detail). They outline that their automated algorithms use a featural (local) processing approach and, therefore, do not rely on holistic information in the way that humans do (see section 2.4 Holistic processing for more detail). Their results showed that human performance was better for all disguises when the faces were familiar, compared to the automated system and suggest that this may in part be due to the lack of holistic processing within the algorithm. However, for unfamiliar stimuli, the results were comparable to humans. It could be argued that the disguises were disrupting the algorithms in a piecemeal way and therefore the threshold for 'matching' becomes more important when there were less features with which to match. However, it is not known if this effect would be replicated when features are changed, rather than obscured.

2.7 Summary

A central question emerged from the general introduction: Does compositing conceal identity through replacing features in known faces? This question can be broken down into two parts: is there a feature saliency hierarchy when target features are replaced, and, how much of the face needs to be replaced in order to conceal identity? Based on these two questions, it is sensible to design two experimental phases to answer them, the first of which is a precursor for the second more comprehensive phase. The first phase of the study will establish if there is a feature saliency hierarchy through individual feature manipulation. The results of the hierarchy will inform the generation of stimuli for the second phase which aims to establish how much of the face needs to be replaced, through compound feature replacement, to conceal identity. This section will outline key methodological considerations, based on previous literature, which will guide the methods of testing as well as the face manipulations required in order to test the role of compositing in concealing identity and answer the questions above. This will then inform the summary of the overall aims and objectives of the study. Alongside this, the literature will be summarized and used to generate and discuss how the compositing manipulations will affect recognition of the target faces. Finally, a summary of the hypothesis is given.

Facial creation

This study aims to evaluate the effect of compositing in concealing identity. There are various ways of compositing and different manipulations that could be used. Therefore, the methodology to be used will be informed by the questions being asked: Is there a feature saliency hierarchy and how much of the face needs to be replaced? With this in mind, it seems that featural manipulations need to occur through the replacement of features of the face. As discussed earlier in the literature review, there are different streams of information gleaned from the face for recognition, including featural, configural (Maurer et al., 2002, Schyns et al., 2002, Tanaka and Sengco, 1997) and texture/contrast information (Burton et al., 2015) (see section 2.4 Holistic processing for more detail). Therefore, it seems sensible to try and control for the other types of information so that only the featural stream is being manipulated. The other streams of information consist of mainly configural and contrast information that make up the holistic processing of faces, as well as meta-data information such as age and sex appearance. Manipulating featural as well as configural and contrast information in the current study would make for a complicated and long experimental process. The most likely compositing process in real world applications involves sampling features and compositing them onto the existing faces in the same position (featural change) rather than just moving the existing features around in order to create a new identity (configural change), or altering the contrast of the face. Therefore, the current study will focus on featural changes only, by replacing features and keeping their position and size congruent with the target face in order to try and minimise any configural changes. However, it is likely that featural changes will inadvertently affect configural relationships, although this may be minimal. The overall contrast of the resulting stimuli will also be kept consistent with the original target face by matching this to the face from which the features have been sampled. Therefore, the counterpart face with which features will be sampled, needs to be carefully matched with the target face.

Feature saliency hierarchy

The literature suggests that there is a feature saliency hierarchy for familiar face recognition (Tanaka and Sengco, 1997, Haig, 1986, Logan et al., 2017) and, as such, part of this study will investigate if there is a saliency hierarchy when features are changed in a target face. This methodology will address the question of whether any features are more important to change than others in order to conceal identity. This question promotes the first experimental phase of the study (Phase 1i) where each feature of the face (eyes, eyebrows, nose, mouth, hair and outline) will be individually replaced in the target face and tested in a face recognition task. The saliency hierarchy results will then be used to generate the stimuli for phase 2. It is expected that this feature saliency hierarchy will remain during compound feature replacement in Phase 2. However, it is not known if the effect of which feature is being replaced will become more or less critical at various points throughout replacement. Given that eyes are considered the most salient from the literature, it could be hypothesised that this would persist throughout all replacement positions. It is also likely that, given this is a familiar face recognition task, preference will be given to the internal features of the face, that are considered more important for familiar face recognition. Therefore, miss rates for the external features are expected to be lower when replaced, than for the internal ones.

Phase 1 will focus on a whole-face context, however, Tanaka and Gauthier (1997) found that recognition was still possible for features shown in isolation when tested following a short familiarisation period with whole faces. Correct recognition was considerably lower than for a full-face context, even when the configuration had been altered. This suggests some kind of whole face interference and dominant role of holistic processing (see section 2.4 Parts and Wholes for a review). It is therefore useful to investigate the role of facial features when shown in isolation as a baseline from which to compare the results from the embedded whole-face condition of the phase 1 experiment. Therefore, a further version of phase 1 (1ii) will show target features in isolation. With this in mind, it is expected that the more salient features, such as the eyes, will yield higher recognition rates when presented in isolation, as they are considered to provide more information. Therefore it is hypothesised that the feature saliency hierarchy results from the embedded Phase 1i experiment, will be inverted for the isolated Phase 1ii version. Given that literature has shown that it is more difficult to recognise facial features in isolation compared to in their congruent whole face, it is likely that recognition rates will be fairly low.

Holistic processing

The second part of the study question asks how much of the face needs to be replaced in order to conceal identity. Hill et al. (2011) suggest that some changes, and the degree to which they're changed, can be made to a face without having any effect on identity whereas other changes will, demonstrating how manipulations can render a face more similar or dissimilar to its veridical self. This aligns with performing a second experiment

(Phase 2) that aims to establish how much of a face can be kept in a composite whilst still concealing identity of the target. Is there a criterion point that indicates how much of the face needs to be changed, in order to create a new identity and conceal the existing one? Features in the target known face will gradually be replaced with unknown ones in a compound manner, to see if any criterion point from old to new identity is found. It is likely that given the use of holistic processing for whole faces, composites will be viewed as new whole faces and the edited parts ignored in favour of perceiving a new identity (see section 2.4 Holistic processing for more detail). However, it is also likely that changing just one feature, as proposed in Phase 1, will leave enough residual face information for the holistic processing to be able to ignore, or moderate for, the edited part. Therefore the results from Phase 1 are not expected to show that identity has been concealed all of the time, but rather that identity is affected by more or less salient features being replaced. Phase 2 will investigate if there is a point at which this is no longer possible when there is too much 'new' information in the face image that results in the image being perceived as a new face through the overriding holistic processing strategy.

If a criterion point for old to new identity is found, it may not only render subsequent featural changes redundant (past the criterion point) but may also rely heavily on which features have been replaced (saliency hierarchy) prior to any criterion point found. An assumption could be made that once the more salient features of the face are replaced, that the criterion point may occur earlier than for if less salient features have been replaced. For this reason, the configuration and order of feature replacement becomes important. Not all orders of features can be tested due to the requirements of stimulus generation and participant numbers. Therefore, the order of feature replacement will be dictated by the feature saliency results obtained from phase 1. To counteract any redundancy effects of feature replacement past a theoretical identity criterion point, the starting point of feature replacement will be rotated to allow for different feature replacement configurations. This will also allow for different configurations of more or less salient features before any criterion point.

Face types

Previous research has shown that memorability, attractiveness and trustworthiness affect our memory for faces (Valentine, 1991, Valentine and Bruce, 1986) (see section 2.5 Facial characteristics for more detail on this), and therefore it was sensible that target faces be rated for these characteristics. Research has shown that distinctive faces are more memorable than average ones, therefore it could be hypothesised that the more the memorable the face is, the more invariant it is to image degradations, or in the case of this study, the replacement of features. It is also important to note that an average face may appear very different to a distinctive face, or any two distinctive faces will also appear very different. Valentine's model also shows that similar faces are clustered together and dissimilar faces further apart on the various dimensions of the face space (Valentine, 1991) (see section 2.5 Face memory and storage). Therefore, any face recognition task can be made easier or more difficult depending on the differences between faces being tested. For this reason it is important that target faces be matched with an unknown counterpart face (unique composite) by taking into account age, contrast/luminance and ethnicity. A large enough target face sample helped to account for differences in the similarity/dissimilarity of target face/unique composite as well as their characteristic averageness/distinctiveness.

Automated face recognition study

It seems likely that, given the nature of creating digital faces for the current study, these images may end up being run through commercially available automated face recognition systems. Therefore, it is important to test the composite stimuli using a sample of automated face recognition systems for performance as well as to compare the results to the human participant data. Due to the proprietary nature of these commercial systems it is impossible to know exactly how these systems are processing the faces. However, academics, who have revealed the underlying processes of their research-based algorithms in published literature, demonstrate that most systems use a method of feature extraction and then use a combination of various processing mechanisms, similar to humans, such as feature shape extraction (featural) and location of the position of features (configural). Research has shown differences between the performance of humans and automated face recognition systems when attempting to recognise faces and it seems that automated systems outperform humans for frontal face images, as will be used in the current study. Therefore, it seems likely that the automated systems will outperform the human participants. In contrast, it is not known if the replacing of features will disrupt automated systems or human performance more. Studies have found that humans outperform machines when recognising disguised faces, although this effect is only found for familiar faces (familiar to humans) and results are comparable for unfamiliar faces. This study will use familiar face stimuli, and therefore it seems likely that humans may outperform the automated systems in this respect. For those systems that use facial features to match novel faces to a learnt face, it seems likely that the replacement of features in the compositing process will perhaps lower the matching score to below that of the systems 'match' threshold figure.

To summarise, both humans and the automated systems may have an advantage in recognising the composite stimuli in the current study and, therefore, it is not known which will perform better: the robust nature of human familiar face recognition may yield higher recognition rates, or the superior performance of automated machines for frontal images may be enough to outperform humans. It is also possible that if the systems use only featural information, that the replacement of one or more features may be sufficient for the system to classify the face as unrecognised. However, if the systems use configural and/or contrast information, this may mitigate the effect of disrupting featural information.

Overall aims and objectives

To test if identity can be concealed by compositing, two questions were initially asked: is there a feature saliency hierarchy when replacing features in a known face and how much of a face needs to be replaced to conceal identity? However, a third question arises, if any particular features have more of an effect on miss rates when they have already been replaced within the compound feature replacement of phase 2?

Aims:

1. To establish if there is a feature saliency hierarchy when target features are replaced and as a baseline, does this hierarchy compare when features are presented in isolation without the interference of a whole face context
2. To establish much of a target face needs to be replaced in order to conceal identity
3. To establish if any features are more important to replace within compound feature replacement

Objectives:

1. Individual features will be replaced in a target face and tested. Differences in miss rates between feature conditions will indicate a feature saliency hierarchy. An isolated feature version will be used to compare to the whole face version
2. Features in the target face will be replaced incrementally and in a compound fashion to see how much of a face needs to be replaced in order to conceal identity.
3. Feature order replacement will be rotated to test for any effects of more or less salient features to be replaced with respect to point 2.

Additionally, two further ancillary studies will be conducted: A target face ratings study, where faces are rated for facial characteristics so that the scores can be used to potentially explain any stimulus specific results from the experimental phases. The second study will take the form of an Automated face recognition system comparison where the composite stimuli will be run through commercially available automated face recognition systems to assess if the compositing technique is sufficient at concealing identity for these systems, which may be adopted in the digital realm.

Hypotheses

1. On the basis that familiar faces are being tested, a feature hierarchy for when features are replaced individually is likely to occur and may mimic results found in previous saliency studies showing the eyes to be the most important and the nose the least. One could hypothesise that the experiment will find differences in miss rates for the effect of which feature has been replaced (changed) in the target face. If the eyes are considered the most salient for familiar face recognition, as shown in the literature, a hypothesis could be made that a target stimulus with the eyes changed would result in higher miss rates than for a target stimulus with the less salient nose feature changed. The isolated feature version is likely to follow the inverse of these results: more salient features such as the eyes will provide more recognition cues than the nose. This suggests some kind of whole face interference and dominant role of holistic processing.
2. Replacing features in a face in a compound fashion is likely to result in a monotonic increase in miss rates. It is not known how much of the target face needs to be replaced in order to conceal identity, but it is likely that some 'criterion' point during feature replacement will be found indicating a shift from familiar target face, to a new identity.
3. It is likely that more salient features, as indicated in phase 1, will result in a significant change in identity at some point through feature replacement (criterion) in comparison to when less salient features have been changed.

Additionally, it is likely that, given the literature on facial characteristics, those faces rated as more memorable will be more invariant to the compositing process proposed in the study, compared to more average faces. Furthermore, it is not known if the automated face recognition study will find that the systems outperform the human data for recognition performance. This is due to research having shown their performance to be better than humans for frontal face images (the pose/view chosen for the current study), but their performance is worse for familiar faces where humans adopt a more robust holistic processing strategy. It is also not known if the replacement of features will disrupt the automated system's performance more than humans. If the systems are purely based on featural shape information, then it seems likely that their performance will be below that of humans. If the systems use additional information such as the configuration of the features and/or the contrast of the image, then they may well outperform humans.

3 Methodology

To test which parts of a face and how much can be used whilst still concealing identity, familiar face targets were manipulated under various conditions to form composite stimuli to be used in a face recognition testing paradigm. Phase 1 tested for a feature saliency hierarchy when presented/manipulated individually, the results of which then informed the generation of the stimuli for phase 2, where features were replaced in an accumulative manner.

Each target face was matched with a counterpart unknown face (Unique composite) and features from the unique composite were sampled and composited onto the target face accordingly to generate six stimulus conditions for each of the experimental phases (Phase 1i and Phase 2: repeated for both lecturer and celebrity faces and Phase 1ii for only celebrity faces = total of five experiments). Furthermore, a separate Ratings study was used to establish stimulus specific information that could support the celebrity results. Additionally, a post-hoc Automated face recognition assessment study using the same stimulus set, was also carried out where the same composite stimuli were run through automated face recognition systems. To summarise;

1. Phase 1i - replaced a single feature of the target with that from an unknown counterpart face (Unique composite) to test if there was a feature saliency hierarchy determined by miss rates.
2. Phase 1ii - showed target features in isolation to provide a baseline of feature hierarchy when features are shown out of a whole-face context, which could be compared to that of the results from Phase 1i.
3. Phase 2 – target features were replaced with those from the unique composite in a compound manner to test *how* much of a face needed to be replaced in order to conceal identity and if any features were more important to be replaced than others.

In this chapter, a more general procedure of obtaining a familiar face set, target recruitment, image processing and matching is outlined. Two pilot studies were carried out to collect the most popular celebrity names for target selection and to assess if the unique composites were convincing as real faces (see Appendix F - Pilot Studies, for more detail). This is followed by a common methodology section that outlines the testing procedure used in the face recognition tasks that was used for all experimental phases, with a separate section on the Ratings task methodology and results.

For Phase 1 and Phase 2, each experimental phase will be described in more detail, including the specific stimulus set outline and any methodological considerations with regards to the experimental design. Results and a discussion will follow the descriptions for each phase (the Automated face recognition study is described last and also adopts this structure).

Ethical approval: All studies received ethical approval to be conducted, see Appendix G for more details [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions].

3.1 Familiar face set

A familiar face set with which to test the compositing manipulations was obtained so that the process of recognition and recall was being tested (see section 2.2 Familiarisation for a review on different familiar face sets).The literature suggested that using a naturally familiar face set is not only more ecologically valid, but also makes for a shorter experimental process with the expectation that the face memories will be well established (Hole and Bourne, 2010, Tong and Nakayama, 1999, Roark et al., 2006, Roark et al., 2003). Two groups of naturally familiar face sets were available for the current study; celebrities (images available online and popular celebrities were likely to be familiar to large group of recognisers) and lecturers (access to staff members in the collaborating institutions and were likely to be familiar to the students they teach).

For the current study, both adult celebrities and non-celebrities (lecturers) were tested (There are known differences in face recognition between adults and children. Therefore, for this study, only adult target faces and participants were used. See Appendix A – Methodological considerations for detail on the development of face processing) and the use of iconic images of the celebrities excluded to avoid the pictorial iconic processing effect (see section 2.2 Different types of familiarity, for a review on this and the use of familiar and personally familiar face stimuli). Further to this, experiments were conducted and repeated using both celebrity and lecturer targets separately so that the experiment could be repeated and the results combined to give a general conclusion that covered a broader range of familiarity that may be more reflective of a real life scenario. It also allowed for the comparison of the two types of familiarity possible. Lecturer targets were recruited from the collaborating institutions and their students formed the participant pool. Celebrity targets come from a much larger pool and, as such, a pilot study (see Appendix F - Pilot Studies, for more detail) was conducted to establish the most popular celebrities, using the same demographic group that formed the participant pool for the experimental phases.

Valentine's model (1991) of a theoretical multi-dimensional face space theorises that more average, or similar, faces are clustered together and more different or distinctive faces are further apart (see section 2.5 Face memory and storage for more detail on this). With this in mind, it was important to be aware of potential differences in results for stimuli depending on the averageness/distinctiveness of the target faces being tested in the current study. Using a large enough target sample size helped to balance out differences by providing a range of faces to be tested. Another subsequent test to collect data on the distinctiveness/memorability of the target faces, as rated by observers (Ratings study) was carried out (see section 3.7 Ratings Study) and the results then used to explain any differences that may have occurred between targets faces in the results sections. It was preferable that the stimulus set for the current study did not contain faces at the very edges of the distinctiveness spectrum. Because of this, target faces were sampled from the predominant ethnic group of the participant pool (in this case, White-European) (see section 2.5 Other-face effects, for more detail on the cross-race effect). All participants were allowed to participate as an assumption was made that they had resided and been exposed to the predominant ethnic population. Target faces and participants were also selected/recruited from all genders and age-groups. However, there were some restrictions on the selection of target faces from older age-groups due to the difficulties in compositing faces with strong age-related changes.

Familiarity thresholds

In the context of this study it was paramount that participants demonstrate their familiarity with the original target so that any poor recognition results for stimuli could be attributed to the manipulations to the face image in the stimulus rather than that the participant was not familiar with the target in the first place (see section 2.2 Familiarisation for a review on how faces are familiar). A Control section was implemented to assess participants' familiarity with the targets. This involved showing participants veridical/unedited images of the targets in a recognition task asking them if the face was familiar. If the familiarity threshold was met, an assumption was made that the participant was familiar enough with the face/identity to be able to recognise it and provide some kind of recall to indicate that it is identity specific retrieval rather than just a familiar response. This enabled the extraction of solely familiar trials for analysis. It is likely that participants had differing levels of face recognition ability, however, this was not tested for as only trials where participants could correctly recognise the face in the control section, were used for analysis (see Appendix A - Methodological considerations, for more detail on face recognition abilities).

Familiarisation period

A proposed minimum natural familiarisation timeframe of one academic semester was used for lecturer targets and their respective student participants for the lecturer versions of the experiments (see section 2.2 Familiarisation for a review of familiarisation periods). There were two practical reasons for this;

1. Given the relatively short time-frame that students were likely to regularly attend university premises and interact with the lecturer targets (two academic semesters per year for undergraduate students) the familiarisation time-frame was

kept short so that students completed the first academic semester, during which familiarisation took place, and were then recruited and tested in the second semester before they left for the summer break (or for 3rd year students, may leave the university). A prediction was made that students were not necessarily taught by the same lecturers throughout their studies and therefore students were recruited during the peak time during which they were taught by a target lecturer.
2. There is a possibility that familiarisation may well have been random and under difficult and varying conditions (viewing distance may be extremely high in some teaching scenarios: e.g. large lecture theatres) therefore a longer familiarisation period may not, in fact, have extended the familiarisation process or rendered the memory encoding more robust due to a familiarity threshold having already been met because of the constraints of the learning environment. This gave additional support to the relatively short familiarisation period of one academic semester.

For the celebrity targets, control could not be gained over the familiarisation period due to celebrity face encoding occurring through participants' choice to view material that contained the target face images (mostly different forms of media). There was also no control over the level of familiarity, similar to the lecturer targets. It was assumed that if the Control section yields a familiar recall that any differences in recognition rates between feature conditions could be attributed to the effects of the stimulus manipulations.

Identity specific recall

This study focused on familiar face recognition with both experiments requiring participants to recall the identities of faces shown to them, and as such this required them to be familiar with the face. How robust this familiar memory needed to be was an open-ended question. As long as they were able to correctly recall the identities of the faces then the familiarity threshold for this particular study has been met. Participants needed to indicate if they had recognised the specific familiar face (target) in some way. Therefore, they were asked to provide the name of the target to indicate this. However, as outlined in Burton et al.'s (1990) Interactive activation and competition (IAC) theoretical model of the processing of faces and names, names are not always accessible or known when a face is recognised as familiar. This model hypothesises that face processing and recognition occurs in a hierarchical node-based manner where faces and names are processed separately and therefore names are not always retrievable upon recognition of a face. Because of this, other researchers have designed their studies to allow for other types of recall in face recognition tasks, such as the recall of identity specific associated semantic information. In a familiar naming match/mismatch task, White (2004) allowed participants to provide a description of the familiar person, indicating recall, as well as the option to type in a name, following on from a key press to indicate familiarity. Davis et al. (2016) also allowed participants to type in a name or associated semantic information that was identity specific. Therefore, participants were also allowed to provide person specific descriptions as an alternative way of describing the identity as these may still have been accessible, even when names were not. For this study it was paramount that participants be able to not only demonstrate familiarity with a face, but to also provide evidence of correctly recalling the specific identity of the individual in the form of face memory associated semantic information.

A Spontaneous naming task was used to test for familiarity of the face stimulus images by implementing three options from which participants could indicate the different types of recognition: unfamiliar, familiar and recall. By allowing a ‘familiar’ option, participants were able to indicate familiarity even when no specific identity could be recalled. For the recall option, providing the name of the identity that they have recalled, or any associated semantic information indicated that the participant had recalled the correct identity. The option to answer ‘familiar’ was implemented to allow for participants to respond in a natural way where faces often appear familiar but semantic information cannot be accessed. This also prevented responses from sliding into the ‘unfamiliar’ option, which would inflate miss rates when in fact the face was not totally unfamiliar. As a corroborative measure, a second task was implemented using a multiple-choice paradigm with name cues to allow for an overt type recognition (Explicit choice), with which the spontaneous naming data can be compared in subsequent analysis. It was important that this was implemented after the Spontaneous naming task so that any occurrence of familiarity that could not be proven with the recall of identity specific information, may be primed with the option to choose from three given names (one name belonging to the target and two distractor names).

3.2 Selection of target faces

Target faces were sourced for the experimental phases for both celebrities and lecturers to form two different versions of the experiments (for Phase 1i and Phase 2, the same celebrity targets were also used for Phase 1ii). These took the form of facial photographs which could then be sampled and composited to form the composite stimuli (see Appendix A – Methodological considerations, for more detail on why photographs were used). The two types of targets were selected in different ways and as follows:

Celebrity targets

The experiments required celebrity face targets that were likely to be familiar to as many participants as possible. Therefore, a Pilot study (Pilot study 1) was conducted to establish twenty-four most popular celebrities using the same demographic group of participants that were used for the main experiments. Participants were asked to name their top ten celebrities in a simple online naming task. From this the top twenty-four most named celebrities were selected as targets. Please see Appendix F - Pilot Studies for more detail on the methods and results of Pilot study 1.

Final celebrity images were sourced and chosen for each of the twenty-four targets. These images were sourced online using internet search engines. Images needed to be of a good resolution and in focus. Editorial images were mostly excluded as these tended to appear unnatural and with specific lighting. Mostly “paparazzi” photographs were chosen as most were in natural outdoor lighting (a few editorial images were used where good paparazzi images were not available). All images contained flash which reduced any lighting inconsistencies and shadows.

Photographs: Images were filtered so that they contained a full face, a frontal view (see Appendix A - Methodological considerations for more detail on why a frontal view was chosen), with the eyes open, mouth closed, neutral expression (see Appendix A - Methodological considerations for why a neutral expression was chosen), no eye-glasses (or to at least have been seen without wearing eye-glasses on a regular basis) and minimal jewellery that could be acceptably edited. Where possible images were chosen with the individual wearing minimal makeup, however this was not always possible as celebrities are mostly photographed at events where a styled appearance has been adopted. More candid paparazzi shots were usually out of focus or from a non-frontal angle and distance, and therefore not appropriate. Where possible, chosen images were those that appeared to have been taken recently to coincide with a contemporary face memory of that person as it was likely that familiar celebrities may have been viewed (familiarised with) over fairly long periods of time. Individuals whose faces had been knowingly altered a great deal over the years through cosmetic enhancement/trauma/disease were excluded from selection. Similarly, individuals who appeared to have a large amount of age-related changes, such as severe wrinkles, were excluded due to the difficulty of compositing faces with large age-related changes to the soft tissues. It was preferable that the stimulus set did not contain very distinctive faces and considering the demographic of the participant pool were predominantly white European, only face stimuli from this ethnic group were selected.

Final targets: Twenty-four celebrity target images were selected, twelve males and twelve females, ~*M*=40 yrs., ~range= 25 – 55 yrs. Images of all of the celebrity targets can be found in Appendix H - Celebrity Targets [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to copyright].

Lecturer targets

To form stimuli for the non-celebrity version of the experiments, photographs of consenting lecturers were recruited from two institutions: Liverpool John Moores University (LJMU) and the University of Central Lancashire (UCLan). Individuals who were teaching at these universities were advertised to via the university email systems with a recruitment email and advert (see Appendix G – Ethics [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]). The recruitment email outlined the criteria for selection, including minimal facial hair, no eye-glasses (or to at least have been seen without wearing eye-glasses on a regular basis) and to ideally teach to groups of students of around fifty or more. Targets needed to be familiar to a large or moderately sized group for students to be suitably familiar. Lecturers needed to have taught a group of students for at least one academic semester to increase the likelihood of familiarity.

Potential lecturer targets were provided with a participant information sheet and consent form (see Appendix G [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]). Once recruited they were asked to complete the consent form and a questionnaire asking for their age, gender identity, ethnicity, nationality and name as well as details of which courses/modules they taught on. Only faces of 'white European' ethnic appearance targets were selected (see Appendix A - Methodological considerations for more detail).

Photographs: Once recruited, a suitable time was arranged to photograph the lecturer's face: this took place in a suitable space (usually the lecturer's office at two sites) where there was mostly neutral lighting and a blank background they could stand in front of. A Canon EOS 700D digital SLR camera with an 18-55mm lens was affixed to a free-standing tripod to photograph the LJMU lecturers. A Nikon D200 digital SLR camera with an 11-70mm lens affixed to a free-standing tripod was used to photograph the UCLan lecturers. The camera was positioned on the tripod at target eye level and at least six feet away from the target lecturer to reduce perspective distortion. Where possible, overhead lights were switched off and the target positioned so that the natural light was cast from the front to minimise shadows. Multiple photographs were taken using various focal lengths, zooming and with and without flash. Auto-bracketing was used to take multiple photos simultaneously with different exposure levels to increase the chance of collecting the clearest photograph. Lecturers were asked to give a neutral expression looking directly at the camera. Images taken using flash photography were selected based on the enhanced focus and detail capturing. This also eliminated any inconsistencies of lighting between targets' surroundings and kept the images congruent with the celebrity images that were also taken using flash photography. Due to the limited availability of lecturer targets for recruitment, the selection criteria were relaxed slightly so that a larger age-range was used and some targets had heavy or short beards. Adjustments were made for these during the stimuli creation (see section 3.5 for more detail).

Final targets: Fifteen lecturer target images, ten females and five males, *M*=42 yrs., range= 27 – 64 years.

3.3 Target images

Lecturer images were uploaded from the cameras' storage cards onto a Lenovo ThinkPad Yoga 14 64-bit laptop (Windows 10) with the celebrity images downloaded from the internet and stored on the same laptop. All target images were stored on the password protected laptop and backed up on a keyed access external hard drive: all equipment was stored in a code accessed research lab situated in a fob accessed building. All target images were opened and processed in Adobe Photoshop CC 2014 with the same settings so that they were consistent across the stimulus set.

The face photographs were first converted to greyscale (see Appendix A – Methodological considerations for more detail on the use of greyscale images) ready for editing. Face portions were selected from the photographs, using the select tool, and pasted onto a neutral grey background so as not to interfere with perception of the face information, at a total image size of 539(W) x 640(H) pixels and a resolution of 144 pixels per inch. The image size was chosen as a portrait orientation so that the image appeared like a normal face image we are accustomed to seeing (e.g. passport) (pixels equated to a physical dimension of 9.51cm x 11.29cm). The size and resolution were also chosen to ensure images could be loaded effectively in the online experimental platform (Qualtrics) without noticeable delay for display whilst keeping the image quality high enough to be able to see face details. Images were saved as jpeg format that compresses the image to reduce the size of the file using a maximum quality level of 12 (minimum compression to maintain the quality of the image) and a progressive format option. Again, this enabled the images to be loaded efficiently by Qualtrics, but without losing too much detail through compression. All target face images were normalized for size with an interpupillary distance of between 130 and 150 pixels to keep the ratio of face to background roughly the same and the relationship between viewer gaze and pupil distance consistent across the stimulus set.

3.4 Unique Composites

In order to test recognition of the familiar faces under various compositing manipulations, an unknown counterpart face was created (Unique Composite) from which features could be sampled and composited onto the target face to generate the stimulus condition images. These unique composites were created so that they could not be identified and their similarity to the target could be controlled for age, sex appearance and, contrast and luminance. Adobe Photoshop CC 2014 was used to composite features using a Lenovo Yoga ThinkPad 14 laptop (Windows 10, 64-bit, using an additional Iiyama 26” ProLite B2483HS monitor). Features were sampled from six different faces for the eyes, eyebrows, nose, mouth, hair and outline. One hundred unique composites were generated: the methodology for generating the unique composites can be found in Appendix B . The compositing technique is outlined in Appendix C .

Once the unique composites had been generated, it was important to establish if the compositing technique was successful in generating plausible faces that would pass as real people. Therefore, a second pilot study (Pilot study 2) asked participants to assess if the unique composites were ‘trustworthy’ or not, as a proxy for whether the face images were considered suspicious or not convincing. Forty-seven unique composites were found to be convincing and formed the pool for matching with target faces for the experiments. Please see Appendix F - Pilot Studies for the full methods and results.

3.5 Matching

As mentioned above in section 3.4 Unique Composites, the unique composites were matched with a target face with consistent age and sex related meta-data as well as matched for the contrast and luminance across the face (see section 2.5 Similarity and dissimilarity of faces for more detail). In order to generate the different stimulus conditions, each target face was matched with a unique composite face so that features could be sampled from the unique composite and composited onto the target face. Only the forty-seven most 'trustworthy' faces, indicated by the results of Pilot study 2, (see Appendix F - Pilot Studies, for results) were used as potential target face counterparts. For the purposes of compositing, the target and unique composites needed to be fairly similar in age and colouring so that the resulting face appeared natural. For example, compositing really dark eyebrows on a very fair face with blonde hair and pale eyelashes would look unnatural. It also made the compositing technique more difficult if there were vast differences in skin texture/contrast. As all images were processed in grey-scale, the eye, hair and skin tones were assessed as either light, medium or dark (including interim shades) as well as assessments of the luminance and contrast of the face image. The scoring was carried out as an observation by the experimenter (subjective). To counter the subjectivity of this observation, an objective method of image processing was adopted simultaneously. Using MathWorks MATLAB 2016a software and the Shine toolbox (Willenbockel et al., 2010) each image's luminance and contrast were calculated:

- Luminance was calculated by averaging (mean) across the luminance of each pixel of the face (mean2 function).
- Contrast was calculated as the standard deviation of the luminance score (std2 function).

The toolbox contained functions for masking off the background of the image (mask function) so that only those pixels that make up the face part of the image were used for the calculations (see Appendix D for the MATLAB scripts).

Once all the target and unique composite faces were assessed, matching of the target face with their counterpart unique composite was carried out using the following criteria:

1. Target images were matched for similar eye, hair and skin-tone observations choosing five of the unique composite faces with the most similar assessments.
2. Image processed luminance and standard deviation scores (contrast) were then used for matching, again choosing five unique composite faces with the closest matching scores.
3. Lastly, the observed luminance and contrast scores were used – these were the most subjective and difficult to observe so were used as the least informative criteria. Again, five unique composite faces with the most similar scores were selected.

From the resulting selections, duplicates across the three different matching methods were selected as possible final matches. The top three highest scoring duplicate images were then put forward for random selection.

Having chosen the top three duplicate unique composite faces, a randomiser was used to allocate one of these faces to their target counterpart face. Given the small pool of unique composite faces to be selected from (n=47), it was likely that more than one target face may include duplicate unique composite faces in their top three matches. Therefore, the randomiser was used to allocate each target face a different unique composite face so that there were no duplicates across the stimulus set within each version (celebrity or lecturer). Several iterations of the randomiser were run until there were no duplicates. A MATLAB script was used to randomise the choices using the 'randi' function (see Appendix D). However, although no duplicate unique composite counterparts were used within the stimulus set *within* each experimental version (celebrity or lecturer), celebrity and lecturer targets could use the same unique composites *between* the two versions as the experiments were separate and participants were recruited so as not to participate in both experiments. Therefore, there was no risk of familiarisation with the unknown unique composite face features that might have occurred. Each matching pair of faces was kept constant for both composite phase experiments (Phase 1i and 2). Once each target face had been matched with a unique composite, any differences in age were adjusted using Adobe Photoshop CC 2014. A large difference in age would result in an unnatural looking face stimulus and could impact on recognition if the ages were incongruent. Due to limited recruitment of lecturer targets, a wider range of ages was used as well as those with some facial hair. In those instances, similar facial hair was added to the matched unique composite so that the mouth and facial outline conditions appeared more congruent with the target. Similarly, ageing wrinkles and age-related tissue changes were added to the unique composite to match more closely with the target using Adobe Photoshop drawing and lighten/darken functions.

3.6 Common Methodology: Phase 1 and 2

All experimental composite phases followed a common methodology and took the same form and structure for the testing procedure, which will be outlined below. The stimulus specific compositing manipulations for Phase 1 and 2 will be described in the relevant experimental phase chapters (see chapters 4.Phase 1 and 5. Phase 2). The automated face recognition study will be described separately in chapter 6. Automated face recognition systems.

Experimental testing was carried out using the online data collection platform, Qualtrics (see Appendix A - Methodological considerations, for a summary of the use of online testing procedures in face perception research). Participants were recruited from Liverpool John Moores University, the University of Central Lancashire and the University of Dundee. For the lecturer version, participants were recruited from the two institutions to which the lecturer targets belonged; Liverpool John Moores University and the University of Central Lancashire. Liverpool John Moores University participants were recruited via department/faculty mailing lists. University of Central Lancashire participants were recruited via the participation system for Psychology students, SONA, and received two course credits for their time. All participants indicated they had normal or corrected to normal vision. Participants were not tested for their face recognition ability prior to testing (see Appendix A – Methodological considerations for more detail).

Tasks

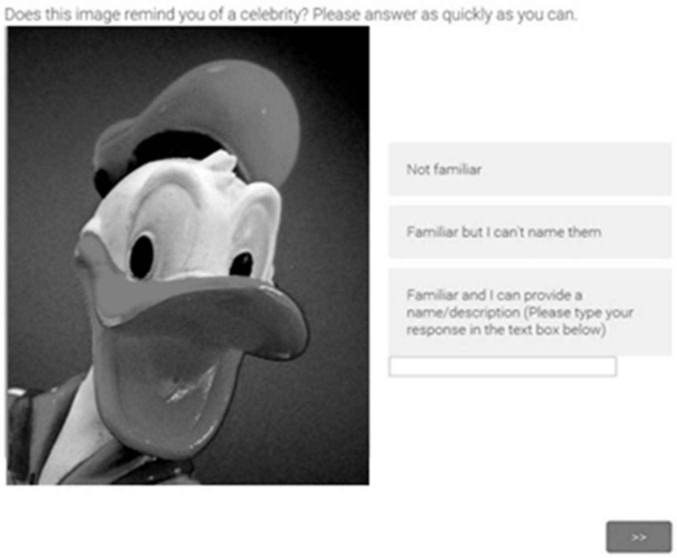
Each experimental phase consisted of a Testing section and a Control section. The Testing section was made up of two tasks:

1. Spontaneous naming Task – stimulus shown, with three familiarity options to choose from
2. Explicit choice Task – stimulus shown, with three name options to choose from

These tasks were designed to accommodate for node based processing of faces and their associated semantic information/names in the brain by allowing participants to respond by indicating their familiarity with the face as well as to further corroborate an identity specific familiarity response (recall) with an input text box of either a name or associated semantic information (Spontaneous naming task). Even if a face is familiar, a name cannot always be retrieved, and therefore the subsequent Explicit choice task provided name cues from which to choose. Each task will now be described in more detail.

Spontaneous naming task: The Spontaneous naming task (see Figure 3.6-1) asked participants to view a face stimulus and requested a forced choice response as to whether the face reminded them of a celebrity or not (or lecturer in the non-celebrity version). Participants were given three options with which to respond:

- Not familiar;
- familiar but I can’t name them;
- familiar and I can name them.



Does this image remind you of a celebrity? Please answer as quickly as you can.

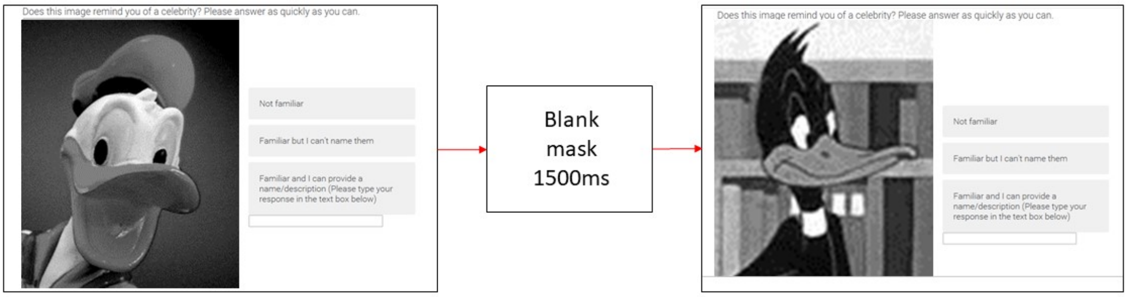
Not familiar
Familiar but I can't name them
Familiar and I can provide a name/description (Please type your response in the text box below)
>>

Figure 3.6-1: Spontaneous naming task of the experiments

An image of Donald Duck is used as an example (see Order of Experiment)

The last response option also required a further text entry for insertion of a name or description of the person the face reminded them of, using the expandable text box provided. Participants were informed at the start that the faces belonged to a pool (celebrity or lecturer). This was given so that the test was made slightly easier to avoid floor results. In a real-life scenario the viewer may not be given any kind of contextual/semantic cue as to which pool/group the face might belong to, and therefore, any miss rate results from the study can be generalised with a more generous effect pattern to reflect the more difficult real-life task.

Participants were requested to respond as quickly as possible so that any identities that the face reminded them of would be noted down, in an attempt to encourage whole-face processing rather than the participant examining the image and attending to specific parts. This procedure also reduced the duration of the experiment so as not to induce fatigue and boredom. A blank screen mask of 1500ms was shown between trials (see Figure 3.6-2).



Does this image remind you of a celebrity? Please answer as quickly as you can.

Not familiar

Familiar but I can't name them

Familiar and I can provide a name/description (Please type your response in the text box below)

Does this image remind you of a celebrity? Please answer as quickly as you can.

Not familiar

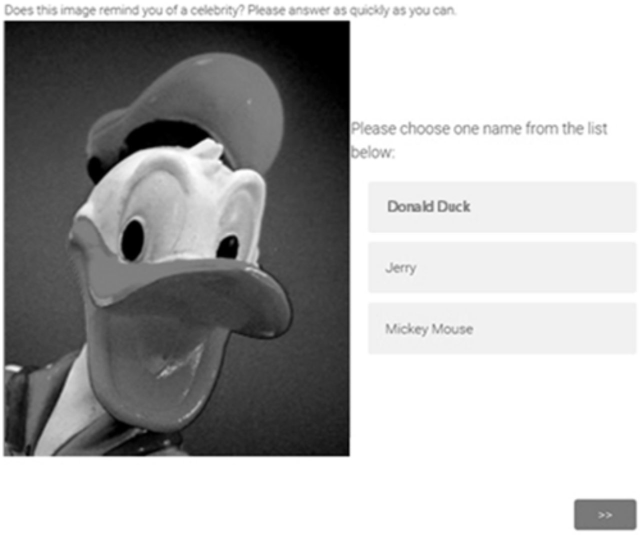
Familiar but I can't name them

Familiar and I can provide a name/description (Please type your response in the text box below)

Figure 3.6-2: Spontaneous naming task order of Presentation.

Participants view a target trial image which stayed on the screen until they clicked to proceed. A subsequent blank screen/mask stayed on screen for 1500ms before the next trial was presented

Explicit choice task: The second task used in the experiment involved an explicit or forced choice test where participants were given three names from which to choose (see Figure 3.6-3: Explicit choice task of the experiments.



Does this image remind you of a celebrity? Please answer as quickly as you can.

Please choose one name from the list below:

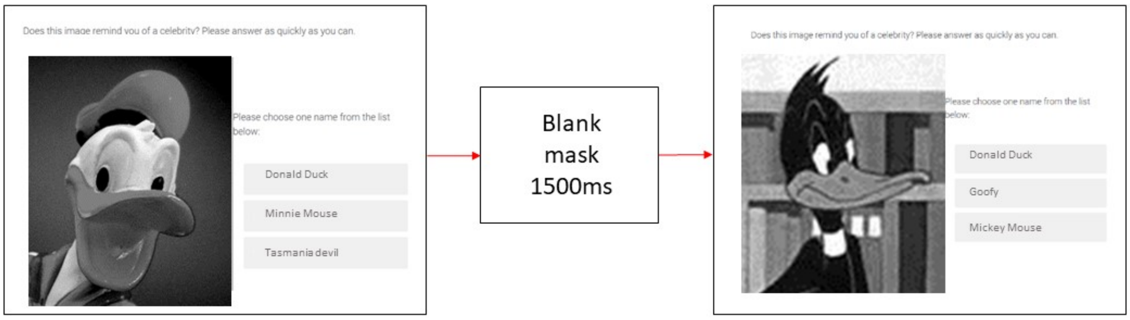
Donald Duck

Jerry

Mickey Mouse

Figure 3.6-3: Explicit choice task of the experiments

One name was the target and two further distractor names randomly sampled from the pool of celebrity names collected in Pilot Study 1, but not including any other target names from the experiment. For the lecturer version, staff names were collected from online university staff profile lists. Therefore, all names used as distractors were not necessarily of individuals of a similar appearance to the target, they were only matched for sex appearance. Each stimulus condition for a target face was randomly assigned distractor names such that no names were repeated across the six conditions for any one target. This was to ensure a balanced distribution of distractor names in case any effects of similarity/dissimilarity or memorability etc. with the target name were to occur. However, this was a familiar face recognition task and as such there was the possibility that participants could choose the correct target name based on a process of elimination of being familiar with the further two distractor names, this issue will be discussed further in the General Discussion (see section 7 General Discussion). Again, participants were requested to respond as quickly as possible with a blank screen mask between trials (see Figure 3.6-4).



Does this image remind you of a celebrity? Please answer as quickly as you can.

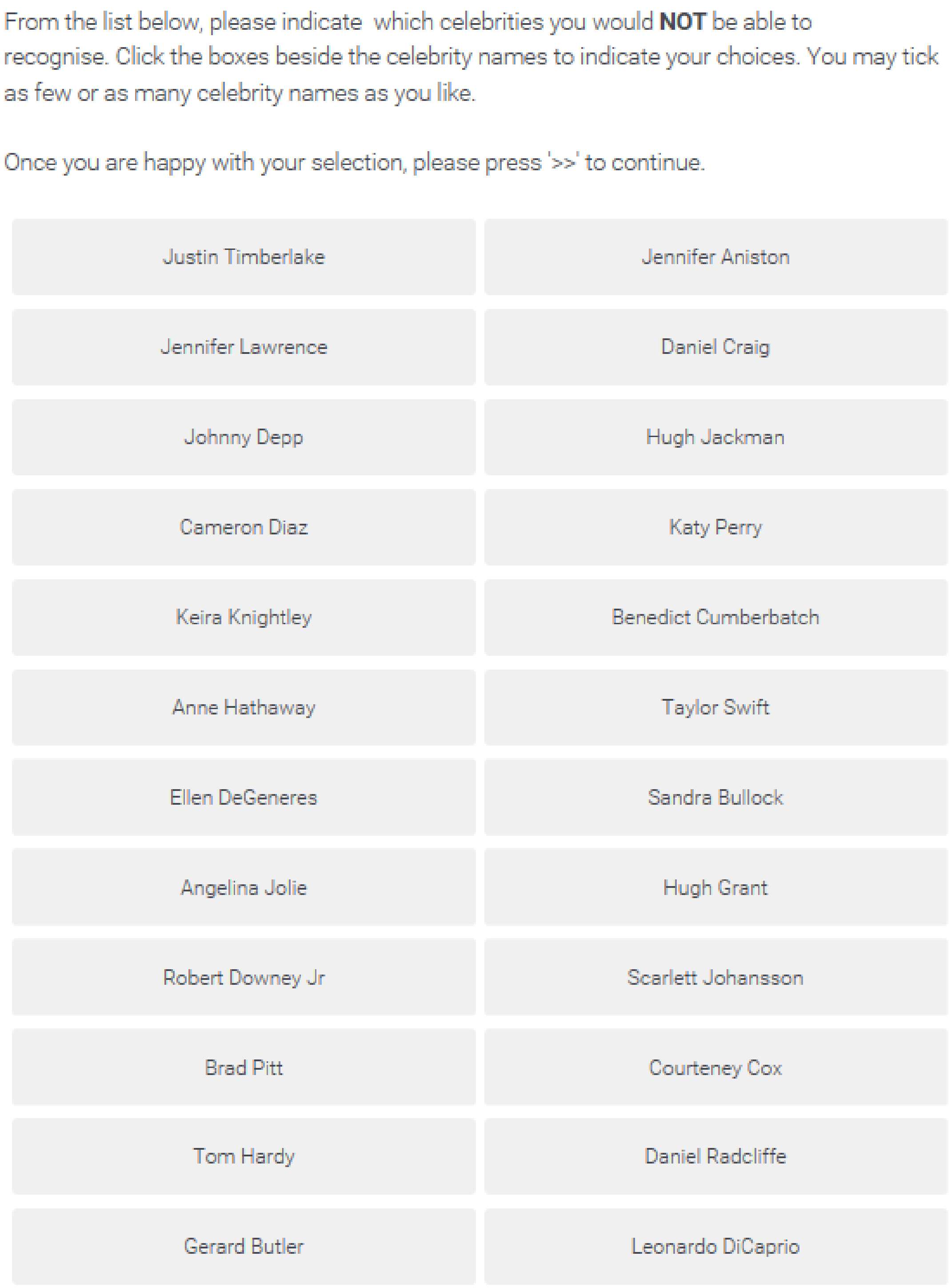
Please choose one name from the list below:

Figure 3.6-4: Explicit choice task order of presentation.

Participants viewed a target trial image which stayed on the screen until they clicked to proceed. A subsequent blank screen/mask stayed on screen for 1500ms before the next trial was presented

Control: The Control section of the experiment was an exact repeat of the Spontaneous naming task in the Testing section above but using the veridical images of the targets (participants were informed that they would now be viewing original images (unmodified) of the targets). The purpose of the Control section was to establish whether the participant was ever, in fact, familiar with the target being tested. The study aimed to investigate *familiar* face recognition and thus it was important to establish familiarity by requesting participants to demonstrate familiarity with the target through identity specific recall. This methodology allowed for the extraction of only familiar trials (so-called ‘valid’ trials) for analysis from which recognition accuracy for the different stimulus conditions could be analysed.

Additionally, a short Exclusion task at the end of the Control section asked participants to indicate (exclude) those target names (identities) with which they were *not* familiar (see Figure 3.6-5), to corroborate results from the Spontaneous naming task in the Control section. All target names were typed and laid out simultaneously on one page where participants could indicate ‘unfamiliarity’ with the target by clicking on the names, these would then turn red to signify selection of an unfamiliar face.



From the list below, please indicate which celebrities you would **NOT** be able to recognise. Click the boxes beside the celebrity names to indicate your choices. You may tick as few or as many celebrity names as you like.

Once you are happy with your selection, please press '>>' to continue.

Justin Timberlake	Jennifer Aniston
Jennifer Lawrence	Daniel Craig
Johnny Depp	Hugh Jackman
Cameron Diaz	Katy Perry
Keira Knightley	Benedict Cumberbatch
Anne Hathaway	Taylor Swift
Ellen DeGeneres	Sandra Bullock
Angelina Jolie	Hugh Grant
Robert Downey Jr	Scarlett Johansson
Brad Pitt	Courteney Cox
Tom Hardy	Daniel Radcliffe
Gerard Butler	Leonardo DiCaprio

Figure 3.6-5: Exclusion task

Participants were asked to select names (identities) that they were NOT familiar with

Presentation of Stimuli

Image stimuli were always placed on the left-hand side of the screen with the answering options to the right. This was to ensure that both stimulus and answer options were visible simultaneously to ease the process of answering and within the framework of Qualtrics capabilities. Due to the nature of online testing, the size at which the face stimuli was viewed could not be controlled because of participant's using different screen sizes and devices for completing the experiment. During instruction they were, however, advised to complete the study on a PC/Laptop rather than a mobile device such a tablet or mobile phone. Therefore, there was no way to establish viewing distance and how much the face stimulus subtended the visual angle. However, to ensure that the whole face image and answering options were visible simultaneously, participants were asked to conduct a screen size assessment at the beginning of the experiment and use the 'plus' and 'minus' zoom functions on their browser to ensure that a whole trial sample was entirely visible on screen.

Face images remained visible until response. A blank white screen (mask) was shown in between trials for 1500ms. Visual adaptation occurs when exposure to an image leaves a residual imprint on the retina that can remain and merge with any new stimulus that is subsequently viewed (Carbon and Leder, 2005a). It is possible that presenting faces in sequence within an experiment may induce some adaptation effects (Leopold et al., 2001). Masks/blank screens and/or fixation crosses between trials and interleaving the stimuli presentation can help to mitigate these effects (Logan et al., 2017, Vesker and Wilson, 2012). Previous face recognition research has adopted this method (White, 2004, Goffaux, 2012, Ramon et al., 2010)), so as to "wipe the slate clean" before a new face image is presented. This study adopted a blank screen/mask of 1500ms between trials in order to reduce/eliminate adaptation effects that might interfere with the perception of a subsequent face stimulus.

Order of Experiment

In this section, the order for the experimental procedure is outlined. This procedure was for all experimental phases.

Introduction: All the experiments began with an introduction, followed by presentation of the Participant information sheet At this point, participants could decide whether or not they wished to take part. A subsequent Consent form was shown and consent was given by participants selecting all the statements and agreeing to take part. In order for participants to be able to withdraw their data (within one week of participating) they were asked to provide a text password in a blank text box, so that the experimenter could then extract the appropriate data, if requested, whilst keeping participation anonymous (see Appendix G for the participant information sheet and consent forms [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]). During this process, Qualtrics also assigned a participant number to each participant.

Questionnaire: Only participants aged 18 yrs. or over were allowed to participate and as such they were required to confirm that they were aged 18 yrs. or over; if they were not, the experiment ended and they were thanked for their interest and time. The next part of the questionnaire asked for age (text box), gender identity (drop down and text box), nationality (self-described text box), ethnic group to which they were most aligned (options given from the British Census) and whether they were in full or part-time education. If they answered yes to the latter, they were asked to provide the year and mode of study as well as the institution. This was to obtain details that would be useful for the lecturer version of the experiment. The other details were requested to see if there were any differences in responses between groups during analysis. All questions were optional (see Appendix E).

Practice: All tasks were preceded by a practice session using images of cartoon characters (like that shown previously in the example figures). This was to illustrate the task required but without providing cues as to the type of images seen in the experiment as well as to avoid any adaptation effects or indications of the identities that might appear in the experiment. Two practice trials were given before each new task.

Tasks: The first part of the experiment consisted of a Testing section where participants completed a Spontaneous naming task, followed by an Explicit choice task, both using the stimulus condition images. A subsequent Control section consisted of a Spontaneous naming task using the veridical images of the targets and an Exclusion task. No feedback was given as to the accuracy of responses.

Testing Section:

1. Spontaneous Naming Task
2. Explicit Choice task

Control Section:

3. Spontaneous Naming Task (veridical images)
4. Exclusion Task

Debrief: At the end of the experiment, a debrief paragraph was given to provide participants with some insight as to the nature of the research. They were also asked whether they received help during the experiment. This question aimed to ensure that participants carried out the study by themselves (see Appendix E).

The experimental order can be seen in Figure 3.6-6.

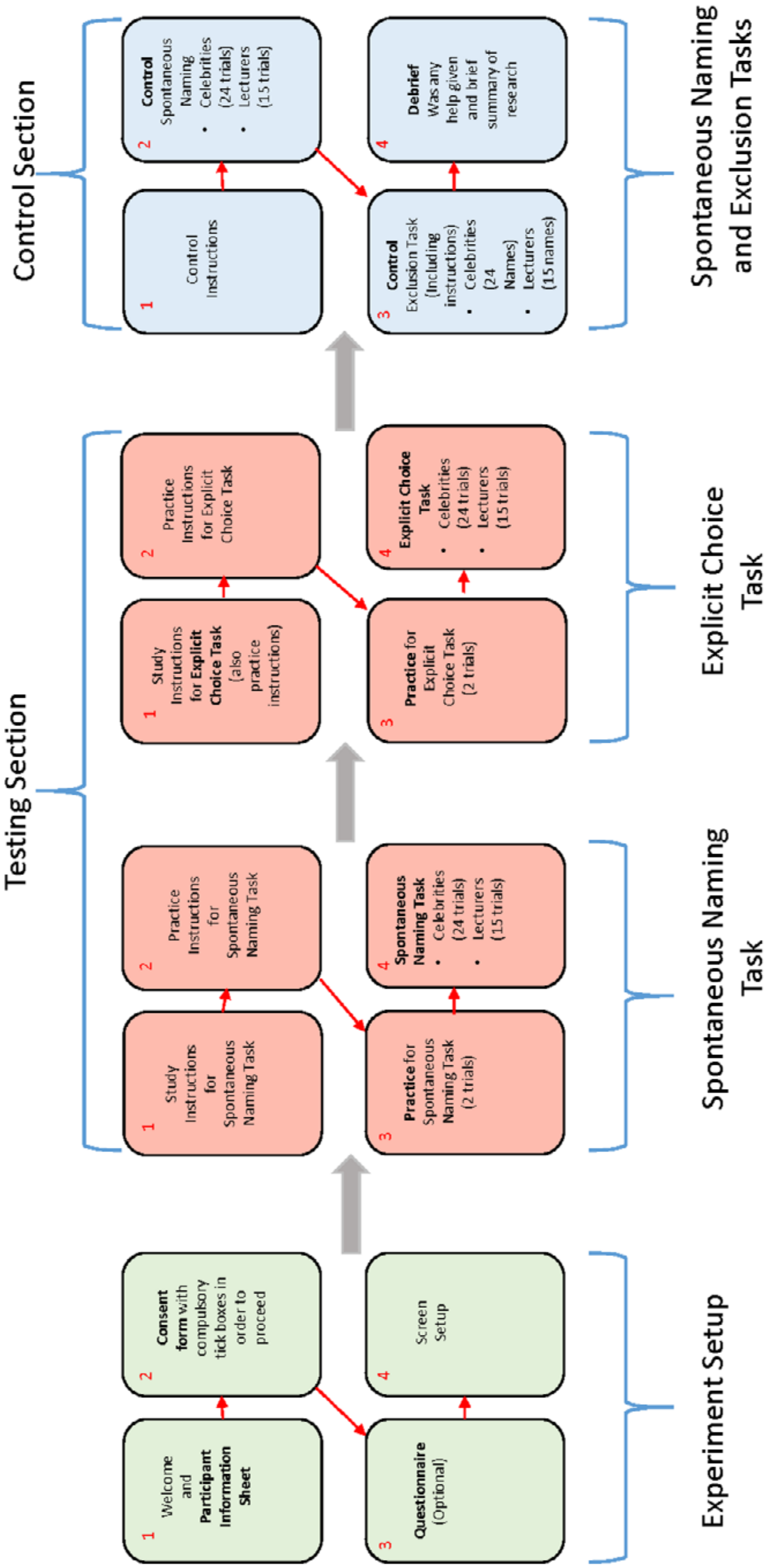

Experiment Setup
1 Welcome and Participant Information Sheet
2 Consent form with compulsory tick boxes in order to proceed
3 Questionnaire (Optional)
4 Screen Setup
Testing Section
Spontaneous Naming Task
1 Study Instructions for Spontaneous Naming Task
2 Practice Instructions for Spontaneous Naming Task
3 Practice for Spontaneous Naming Task (2 trials)
4 Spontaneous Naming Task
• Celebrities (24 trials)
• Lecturers (15 trials)
Explicit Choice Task
1 Study Instructions for Explicit Choice Task (also practice instructions)
2 Practice Instructions for Explicit Choice Task
3 Practice for Explicit Choice Task (2 trials)
4 Explicit Choice Task
• Celebrities (24 trials)
• Lecturers (15 trials)
Control Section
Spontaneous Naming and Exclusion Tasks
1 Control Instructions
2 Control Spontaneous Naming
• Celebrities (24 trials)
• Lecturers (15 trials)
3 Control Exclusion Task (Including instructions)
• Celebrities (24 Names)
• Lecturers (15 names)
4 Debrief Was any help given and brief summary of research

Figure 3.6-6: Experimental order

Order of Sections and tasks in the Experiments

Participants only ever saw a target (in its respective feature condition) once in each of the tasks in the Testing Section. Repetition priming, in the context of face recognition studies, is the activation of identity associated nodes through the prior exposure of a related face image (Ellis et al., 1990, Burton et al., 1990). As a recognised occurrence in face recognition studies (Bindemann et al., 2007, Bourne et al., 2009, Goshen-Gottstein and Ganel, 2000, Steede and Hole, 2006), this phenomenon was eliminated by only showing target faces twice throughout the Testing section of the experiment: once each during the Spontaneous naming and Explicit choice tasks. The Spontaneous naming task was perceived to provide the most useful information as to how the stimulus was performing with regards to recognition and recall and thus was placed prior to the less informative Explicit choice section that could have carried over some repetition priming effects. However, a Control section was also necessary, as mentioned earlier, to assess for familiarity with the target image by showing veridical face images and in turn repeating exposure to target faces. It was important to ensure that the Testing section occurred prior to this Control section so that any repetition priming that might occur through multiple exposure, occurs after the test section (Testing also occurred before the Control section for the obvious reason of not showing a target before it has been seen its respective testing conditions). It is possible, therefore, that the Control section could have been facilitated by the prior exposure to a condition stimulus of the target, as well as through the semantic cue of 'celebrity' or 'lecturer' and/or through the Explicit choice task. However, the benefit of the Control section was considered to far out way such a risk.

Qualtrics: Within Qualtrics, each target's six stimulus condition images were pooled, from which Qualtrics randomly chose one stimulus to show (see Appendix E - Testing). Therefore, participants saw an unbalanced number of stimuli for each condition across the experiment, although participants would always see each of the twenty-four targets once for the celebrity version. LJMU participants saw each of the nine targets once, however, some lecturers were from incongruent departments. UCLan participants saw each of the six targets once and all lecturers were from a congruent department to the participant pool. However, as there was no predictor of valid familiar trials per participant, even if control had been obtained, there would still be an uneven sample size of valid trials per condition per participant to analyse.

For the Testing Section, the order of target trial presentation (identities) was randomised in the Spontaneous naming task. The subsequent Explicit choice task used the same stimulus conditions as in the Spontaneous naming task, again randomizing the order of trial presentation. Name choices were also randomised for presentation in the Explicit choice task. Within the Control section, the Spontaneous naming task was set up with only one option of stimulus: the veridical image. Again, the order of trial presentation was randomised. The order of name boxes in the Exclusion task remained the same. It should be noted that presentation order by identity was different between the Spontaneous naming and Explicit choice tasks for a participant.

Participants saw a total of 73 stimulus trials for the celebrity versions and 37 trials for the LJMU lecturer version and 19 trials for the UCLan lecturer version (see Table 3.6-1).

Table 3.6-1: Total trials seen for the different versions of the experiments

		Version (Trials)		
Section	Task	celebrity	lecturer (LJMU (L); UCLan(R)	
Testing	Spontaneous Naming Task	24	9	6
	Explicit Choice Task	24	9	6
Control	Spontaneous Naming Task	24	9	6
	Exclusion Task	24 Names (1 Trial)	9 Names (1 Trial)	6 Names (1 Trial)
	Total	73	37	19

Data were collected mostly simultaneously for Phase 1i and 1ii celebrity versions with Qualtrics setup to randomly assign participants to either version in a balanced manner using the same experimental link. Quotas were assigned to collect a maximum of 100 participants (power analyses suggested 90 participants) per phase so that if the quotas were met, participants would be blocked from participating. Quotas were met for phase 1i before phase 1ii and therefore results for phase 1i were analysed first.

The results of phase 1i (celebrity) were then used to generate the stimuli for phase 2 (celebrity) and once complete, phase 2 was made available as a separate experiment in Qualtrics. Phase 1i and phase 2 (for each lecturer versions; UCLan and LJMU) formed one experimental link each in Qualtrics where participants were randomly assigned to either phase 1i or phase 2 and were run simultaneously to phase 2 (celebrity). Again, quotas were set up with a maximum of 100 participants per phase. Participants were only permitted to complete one experiment. All responses were recorded by Qualtrics and stored on their server. Additionally, reaction times were recorded for the total duration of the experiment as well as for first click timings (when participants first clicked on the page, most likely to answer) and page submit timings (how long the trial stayed on screen) as well as stimulus presentation order. However, due to the nature of online data collection and differences in internet speed, trial response times were considered unreliable for analysis.

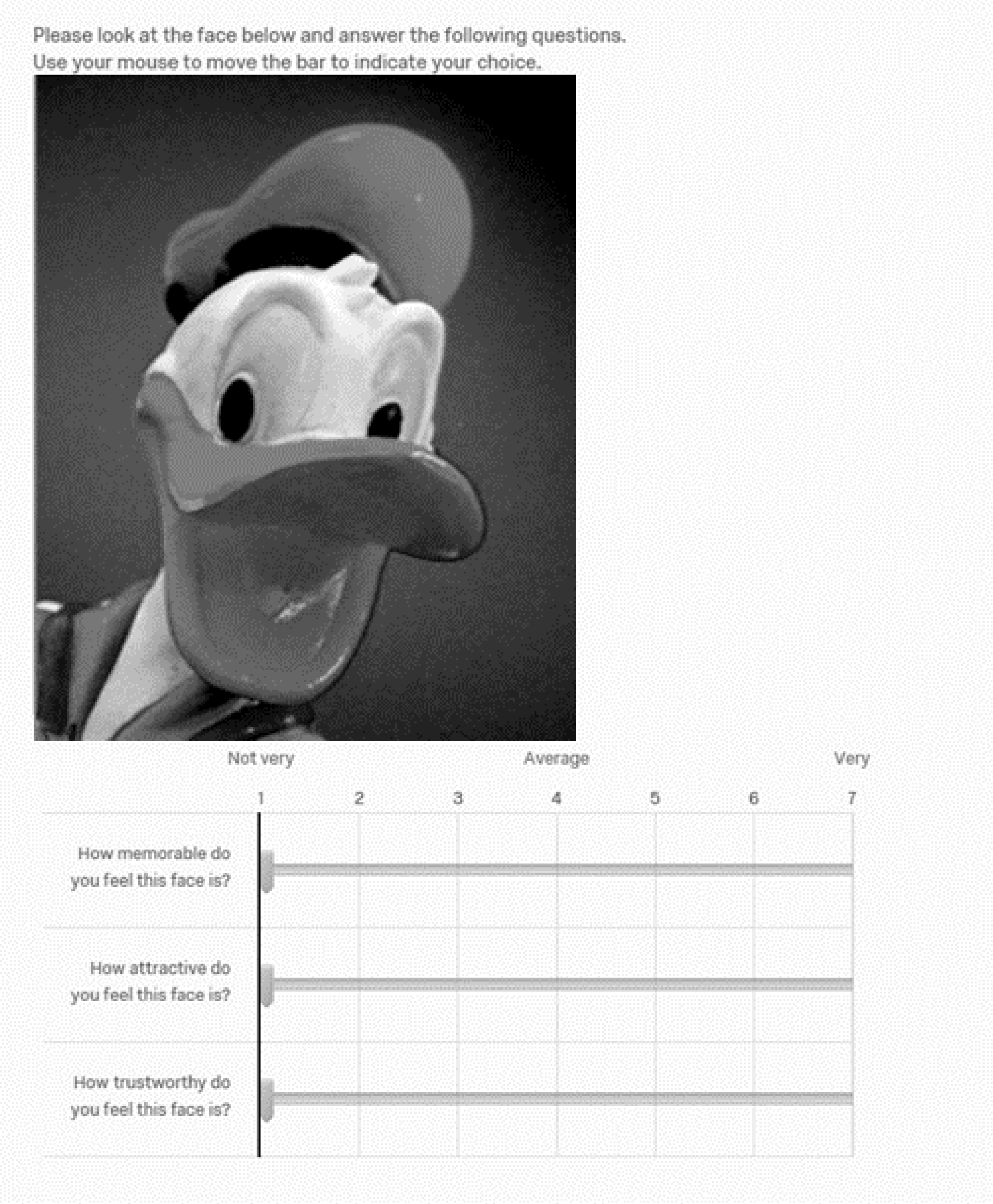
Liverpool John Moores University participants were sent the experimental link to complete the study online using their own devices (preference for PCs/Laptops as stated in the experimental instructions). University of Central Lancashire participants were only given the experimental link via an online participant system (SONA) once they had signed up to participate for the online version. They then completed the experiment online in the same manner as Liverpool John Moores University participants, or, they were given a timeslot to attend one of the experimental laboratories based in the Psychology school. During attendance the experimental link was loaded on the lab-based PC, via the internet, for them to complete. Only the participant and experimenter were present in the laboratory during the study.

The celebrity version of the experiments for all phases took, on average, approximately 24 minutes, and approximately 13 minutes for the LJMU version of the experiments and 10.5 minutes for the UCLan versions.

3.7 Ratings Study

In a separate ancillary study, the twenty-four celebrity target images to be used in the subsequent experiments were assessed for three characteristics: memorability, attractiveness and trustworthiness (see section 2.5 Facial characteristics for a review of characteristic ratings). Each characteristic was rated using a seven-point Likert scale; 1 indicating the least evidence for that characteristic (1 = ’not very’), 4 indicating average evidence (4 = ’average’), and 7 indicating the most (7 = ’very’). The three Likert scales were shown simultaneously below the target image (see Figure 3.7-1).

Procedure: Participants were shown a participant information sheet to read and then a consent form where they indicated their consent by clicking all the statements. A subsequent questionnaire was presented (same as for the experiments above). Participants first completed a practice session using cartoon images (two trials) and then proceeded to complete the main part of the experiment. In the main part they were asked to look at each of the twenty-four veridical celebrity target faces, one-by-one, and rate them for the three characteristics (see Figure 3.7-1).



Please look at the face below and answer the following questions.
Use your mouse to move the bar to indicate your choice.

How memorable do you feel this face is?

How attractive do you feel this face is?

How trustworthy do you feel this face is?

Figure 3.7-1: Celebrity Ratings task

Screenshot of a typical stimulus presentation for the celebrity Ratings task. The three characteristic Likert scales were presented simultaneously below the stimulus.

Familiarity with the targets was established using a subsequent Control section in the study consisting of a Spontaneous Naming task and Exclusion task, the same as that used for Phase 1i, 1ii and 2 (see Figure 3.6-1). Participants were then asked if they had received help or not and were provided with a debrief sheet (same as for the experiments above). No feedback was given as to the accuracy of responses and an inter-trial blank screen/mask was shown for 1500ms. The study was not repeated, as intended, using the lecturer target faces due to the limitations in recruiting participants familiar with the targets.

Participants: Fifteen participants (male = 2, female = 13), age: *M* = 24.93 yrs. (Range: 18-38 yrs.) [not declared = 1] completed the Lab-based study. Participants were undergraduate students from the Psychology department of the University of Central Lancashire, recruited via the online participation system, SONA (For participant information sheets and consent form, see Appendix G [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]). All participants indicated that they had normal or corrected to normal vision and were aged 18yrs or older.

Results

Data screening and scoring: A total of 15 responses per item (n = 24) were collected (total trials = 360). Scores were collected as a rating on a Likert scale from 1-7 for the three characteristic ratings, and familiarity was scored as those trials where the participant correctly named the target in a Spontaneous naming task, as a percentage of the total trials shown. A check for missing data was made, of which no cases were found.

Table 3.7-1: Celebrity ratings by characteristics

Characteristics	Mean	Standard Deviation	Minimum	Maximum
memorability	5.4	0.5	4.27	6.40
trustworthiness	4.0	0.7	2.93	5.53
attractiveness	4.6	0.7	3.00	5.47

Mean scores were collapsed across items (based on a Likert scale of 1-7) for the three tested characteristics

Descriptive analysis: Table 3.7-1 shows the mean memorability characteristic was the highest (*M* = 5.4) and much higher than the other two, followed by attractiveness (*M* = 4.6) and trustworthiness (*M* = 4.0). The memorability characteristic was rated most consistently (*SD* = 0.5) than attractiveness and trustworthiness *(SD* = 0.7).

Correlations analysis: A correlational analysis was used to assess if familiarity with the target affected perception of the three characteristics tested across the twenty-four target faces (see Table 3.7-2).

Table 3.7-2: Celebrity ratings - Correlation between Familiarity and characteristic ratings

		attractiveness	memorability	trustworthiness
Familiarity	Pearson Corr.	0.1	0.7	<.01
	Sig. (2-tailed)	0.66	<.01	0.86
	N	24	24	24
attractiveness	Pearson Corr.		0.3	0.4
	Sig. (2-tailed)		0.21	0.09
	N		24	24
memorability	Pearson Corr.			0.1
	Sig. (2-tailed)			0.61
	N			24

Results showed a significant positive correlation between familiarity (*M* = 68.9%, *SD* = 17.6%) and memorability (*M* = 5.4, *SD* = 0.5) (Pearson's *r*(22) = .69, *p* < .001); that is, the more familiar someone is, the more memorable their face. There was no significant correlation between familiarity and trustworthiness *(M* = 4.1, *SD* = 0.7) (*r*(22) = -.04, *p* = .86) nor between familiarity and attractiveness (*M* = 4.6, *SD* = 0.7) (*r*(22) = .09, *p* = .66). There were also no significant correlations between the three characteristic ratings (*p* > .08) although approaching between attractiveness and trustworthiness (*p* = .09).

Discussion: Memorability was rated the most consistently compared to the other two characteristics, as indicated by the standard deviation results. This suggests that there was more variation across the participant group for perceptions of attractiveness and trustworthiness from faces. It is possible that memorability was interpreted as more of a physical characteristic whereas attractiveness and trustworthiness were more personal subjective observations of visually derived semantics. Memorability was also the only characteristic affected by whether participants were familiar with the face or not with a positive correlation between the two. It is possible that a sense of familiarity with the face in turn indicated to the participant that they were memorable as they already had a memory for that face. This finding is known in the literature and was expected (Valentine, 1991, Valentine and Bruce, 1986). There was an approaching significant correlation between attractiveness and trustworthiness, in line with previous literature showing perceived attractiveness is modulated by personality preferences (Botwin et al., 1997, Little et al., 2006). Overall, the characteristic ratings were used to describe stimuli for the experiment in a specific way, which were then used as covariates in the analysis of the experimental phase results to see if there were any stimulus specific effects that may have affected miss rates (see sections 4.3 Results (phase 1) and 5.3 Results (phase 2)).

3.8 Design Summary

Two main experiments were conducted (Phase 1 and Phase 2), with Phase 1 split into two parts: Phase 1i, Phase 1ii and Phase 2:

- Phase 1i - individual features were sampled from the counterpart unique composite face and then composited onto the target face
- Phase 1ii -individual features from the target face were shown in isolation.
- Phase 2 - showed the target face with incrementally replaced features (compound) sampled from the counterpart unique composite, using the feature hierarchy results from Phase 1i (celebrity) for order of replacement

Phase 1i and 2 were tested using both celebrity and lecturer target faces. Phase 1ii was only tested on celebrity target faces.

Design: Each experimental phase included six different stimulus conditions per target face. Participants saw only one trial per target face. There were 24 trials for the celebrity versions, 15 for the lecturer version. One stimulus image was sampled from the six available conditions per target (Phase 1: eyes, eyebrows, nose, mouth, hair and outline; Phase 2: replacements 1-6). The stimulus selections for each target were kept constant for all of the Testing tasks in the experiment, per participant. For example, Participant 1 may have seen target one in condition A (eyes changed), target two in condition C (nose changed), target three in condition D (mouth changed) etc. Participant 1 would carry out all Testing tasks using this same stimulus selection. An example of a configuration of stimuli seen by a participant can be found in Appendix I [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to copyright].

Sample size: The aim was to be able to detect practically useful effects, particularly those that have an impact in the real world, and therefore the experiment was designed with sufficient power to be able to detect a medium-large effect size, should one exist. Power analyses using the statistical package, G*Power 3 (http://www.gpower.hhu.de/en.html) indicated that approximately 90 participants (15 participants per condition) were required to be able to detect a large effect size (ANOVA, omnibus, one way and six groups with a large effect size f = 0.4, power = .8 and alpha = .05) based on correct recognition (as assessed by correct names). This estimation was for each phase of the experiments.

Analysis

All response data was downloaded from Qualtrics and stored as Microsoft Excel files. Data was organised, collated with basic analysis in Excel followed by the use of IBM SPSS 23/24 for inferential statistics and generation of tables and graphs. All SPSS output for analysis can be found in Appendix J [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]. Only trials where participants could prove that they were familiar with the target face, through a Spontaneous naming task using veridical images in a Control section, were extracted. Miss rates were calculated and used for analysis: these contained “unfamiliar” responses, in addition to inadequate (no indication they recalled the correct identity) or incorrect naming (this indicated that the stimulus represented a different identity). In order to test for the effect of replacing a feature in the target face, an analysis of variance (ANOVA) was carried out for all phases, varying between within and between-subjects Univariate one way or Repeated Measures ANOVA depending on the specific analysis being done. Condition means were compared to each other in post-hoc tests to assess for feature saliency hierarchies as well as any correlations between conditions. Phase 2 results allowed for the analysis of the overall effect of replacement as well as for unpacking the impact of which features had been replaced throughout the different replacement conditions. Post-hoc tests were used for analysing any subsets of levels that might provide an insight as to whether there was a criterion point between old and new identity. Cohen’s D was used to measure effect sizes between condition levels. Results from the Ratings study (see section 3.7 Ratings Study) were also used as covariates in further ANOVA analyses of the experimental phases to test if stimulus characteristics affected condition results.

4 Phase 1

4.1 Introduction

The first experiment aimed to establish if there was a facial feature saliency hierarchy when individual features were replaced on a target face. Celebrity and lecturer target faces had features replaced with those from their paired unique composite face counterparts, to produce the composite stimulus set., which will be described in further detail below in the Method section (see 4.2 Method). The testing methodology for this stimulus set is outlined in the Common Methodology section (see 3.6 Common Methodology: Phase 1 and 2) where data was collected online in a face recognition task. The feature hierarchy results of Phase 1 then informed the stimulus generation for Phase 2 (see section 5.2 Method).

4.2 Method

Materials: Unknown features from the unique composite faces were composited onto the target face one feature at a time to produce six different stimulus images for one target face (six conditions: eyes, eyebrows, nose, mouth, hair and facial outline). These features were composited using Adobe Photoshop CC 2014 and aligned and resized to fit and maintain the configuration of the target face so that only feature shape and texture were being changed and therefore tested, see Figure 4.2-1. For details of the compositing technique, see Appendix C .

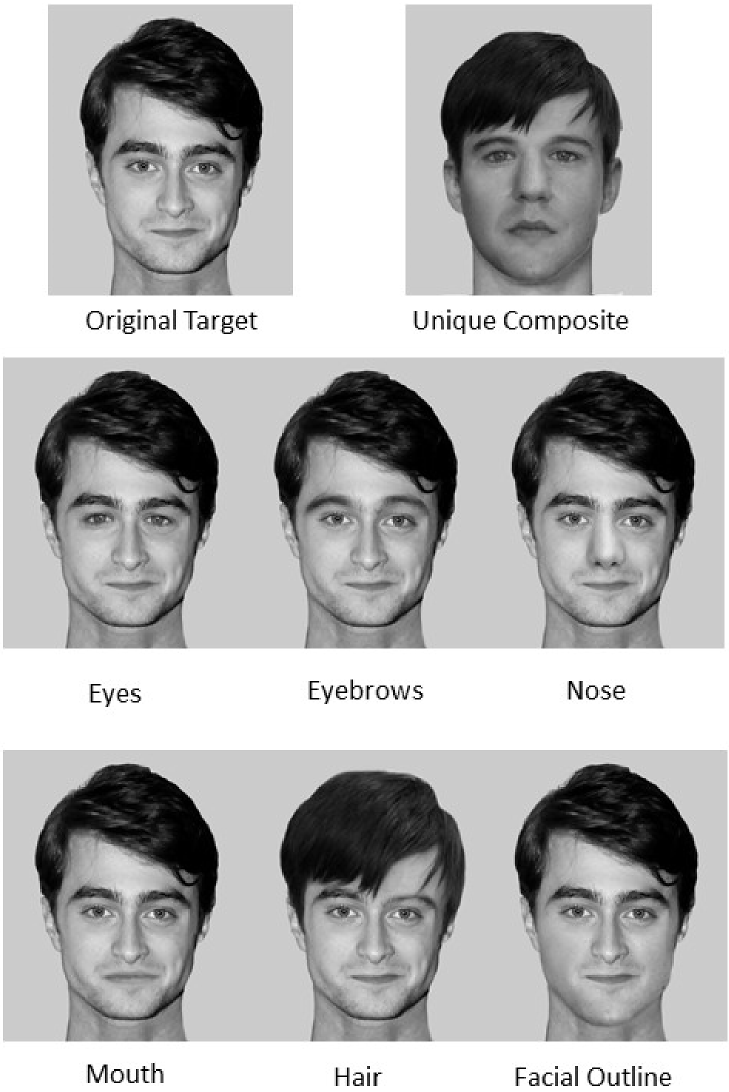


Figure 4.2-1: Phase 1i example stimulus set indicating individual feature replacement

The six feature conditions (middle and bottom rows) generated by sampling individual features from the unfamiliar unique composite (top right) and compositing them onto the target face (top left)

4.3 Results

Data screening and scoring: Only trials where participants could indicate that they were familiar with the target celebrity, in the Control section, through correct spontaneous naming of the veridical images, were used for analysis. Participant responses were scored for accuracy for each target composite in the Control section. For the Testing trials, a value of “1” was assigned when participants gave the correct name. A value of “0” was assigned for invalid trials where participants responded, “unfamiliar”, or “familiar” or they gave the incorrect name. Miss rates were calculated as those trials where participants indicated that the face was not familiar in the Spontaneous naming task (Testing section) or where they thought the face was familiar but gave the wrong name (incorrect naming) as a percentage of the familiar (valid) trials (“familiar” responses were not included in this analysis).

Celebrity version (1i)

Participants: One hundred participants aged 18 or over (age: *M* = 29.1 [age not declared = 5] range = 18-64 yrs.) completed the online study (males n = 30, females n = 68, other n = 1, not declared n = 1).

Data: A check was made for missing data, of which no cases were found. This resulted in mean familiar (valid) trials of 14.5 per stimulus (total trials seen: *M* = 16.7). This meant that 87.2% of all trials were familiar and, therefore, valid for analysis (total valid trials = 2089 (*M* = 348.2 per feature)).

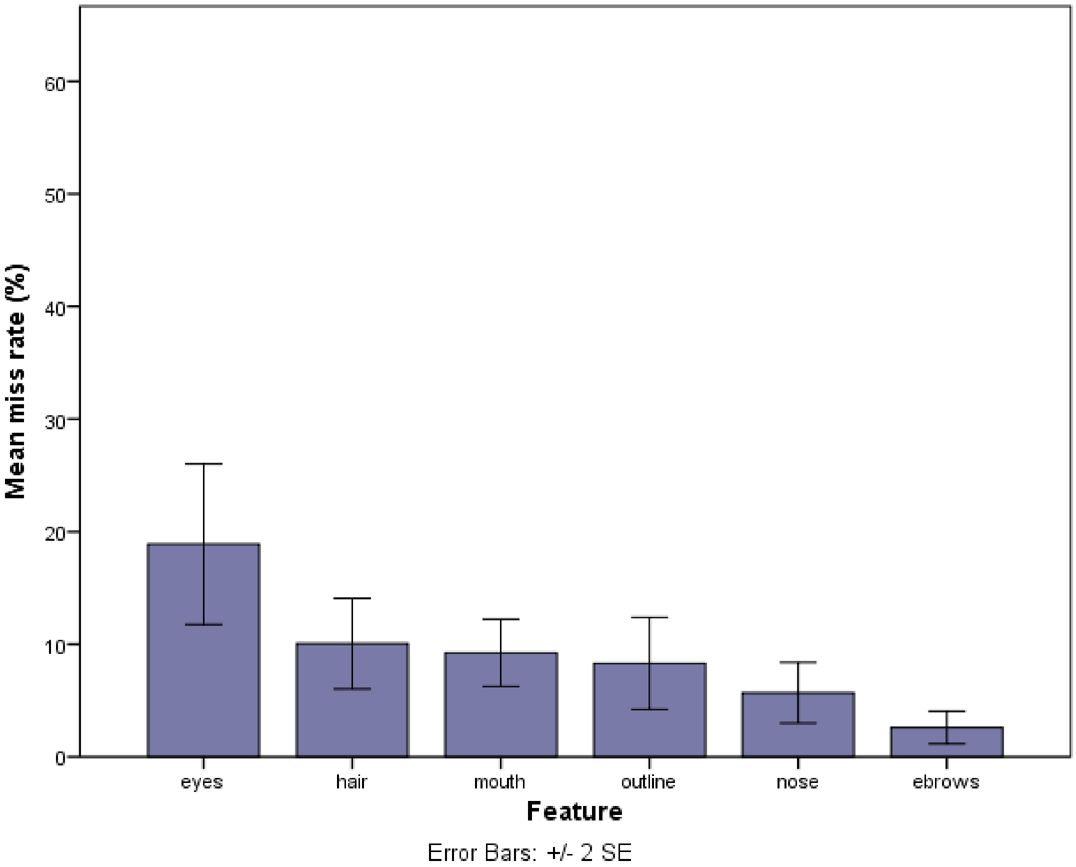


Figure 4.3-1: Phase 1i (celebrity) – Mean miss in percentage by feature.

Collapsed across 24 items. All mean values contain a minimum of 329 observations.

Descriptive analysis: Figure 4.3-1 shows overall mean miss (%) scores were low (<18.9, *M* = 9.1%, *SD* = 11.1 %). For the factor of feature, a clear order of miss rates was observed with eyes (E) (*M* = 18.9%, *SD* = 17.5%) yielding the highest and eyebrows (EB) (*M* = 2.6%, *SD* = 3.5%) by far the lowest overall. Hair (H) was next highest (*M* = 10.1%, *SD* = 9.9%), similar to mouth (M) (*M* = 9.2%, *SD* = 7.3%), and outline (O) (*M* = 8.3%, *SD* = 10.0%). Nose (N) (*M* = 5.7%, *SD* = 6.6%) was intermediate to outline and eyebrows. Eyes' miss rate is almost double the miss rate of the next highest feature (H). In support of the feature saliency hierarchy, the number of items unaffected by the manipulation is the inverse of the increase seen in the mean miss rates (E = 3, H = 7, M = 6, O = 10, N = 11, EB = 15).

Analysis of Variance: To test for the overall effect of facial characteristic ratings on miss rates, an analysis of variance (ANOVA) was carried out. For the full model, an items analysis (n = 24) (female = 12, male = 12) mixed-factor Repeated Measures (RM) ANOVA for the within-subjects factor of feature (E,EB,N,O,M,H), between-subjects factor of target sex (female, male) and covariate of memorability (scores taken from the Ratings Study) was used. This showed an overall effect of memorability (*F*(1,21) = 7.35, *p* = .013, η_p^2 = .26) and an approaching significant interaction with feature (*F*(2.83,59.36) = 2.58, *p* = .065, η_p^2 = .11). Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated (X^2(14) = 40.85, *p* < .001) and therefore, a Greenhouse-Geisser estimate was used (ε = .57). A scatterplot was plotted to establish how memorability affected miss rates. This indicated a negative correlation between the mean miss rate and memorability (Field, 2016) (see below for further analysis for this covariate). The analysis was repeated separately for the covariates of attractiveness and trustworthiness. Results showed no significant effect of attractiveness (*F*(1,21) = 0.16, *p* = .69, η_p^2 = .01) or interaction with feature (*F*(2.68, 56.32) = 0.26, *p* = .83, η_p^2 = .01). There was also no effect of trustworthiness (*F*(1,21) = 0.00, *p* =.96, η_p^2 = .00) or interaction with feature (*F*(2.71,56.87) = 0.37, *p* = .75, η_p^2 = .02).

To analyse the overall significant effect of memorability, mean miss scores were collapsed across feature to give an overall miss rate for each target. Using a correlations analysis, these scores were then analysed against the memorability covariate scores to see how memorability of a target face affected its miss rates. Results showed a significant negative correlation (Pearson's *r*(22) = -.42, *p* = .039). This means that the higher the memorability score (more memorable), the lower the miss rate overall.

Simple main effects: To further investigate the approaching significant interaction between feature and memorability, simple main effects were used to assess the effect of memorability on each feature. Results showed a significant effect for the outline feature (*F*(1,22) = 5.24, *p* = .032, η_p^2 = .19). Approaching significant effects were found for eyes (*F*(1,22) = 3.74, *p* = .07, η_p^2 = .15) and hair (*F*(1,22) = 3.38, *p* = .08, η_p^2 = .13). All other features were not affected by memorability (*p* > .62). To understand how target memorability affected the composites where the outline had been changed, miss rates for outline were correlated against the memorability scores. Results showed a significant negative correlation (*r*(22) = -.44, *p* = .032): the higher the memorability score, the lower the miss rate for outline.

To analyse the effect of feature and target sex on miss rates, the covariate scores were removed from analysis (in an attempt to increase statistical power). RM ANOVA revealed a significant effect of feature (*F(*2.72, 59.86) = 8.35, *p* < .001, η_p^2 = .28, $X^2(14)$ = 46.78, *p*<.001, ε = .54) on miss rates, meaning there was a difference in miss rates between features, overall. There was no significant interaction between target sex and feature (*F(*2.72,59.86) = 0.6, *p* = .60, η_p^2 = .03). The between subjects results show no main effect of target sex (*F*(1,22) = 0.69, *p* = .41, η_p^2 = .03). Both these results showed that target sex did not affect miss rates and the effect of feature replacement was not affected differently by sex.

To further investigate the overall effect of feature, post-hoc pairwise comparisons (Bonferroni corrected) were conducted to establish if any features had statistically different effects on miss rates compared to others. Results revealed eyes (highest miss rate) were significantly greater than eyebrows (*p* = .003), nose (*p* = .028) and outline (*p* = .032). Eyebrows (lowest miss rate) were also significantly different from mouth (*p* = .003) and hair (*p* = .022). Results for each feature were also compared to each other to establish if any features followed a similar pattern of effects (positive or negative) on miss rates to others. Results revealed a significant positive correlation between eyes and outline (Pearson's *r*(22) = .51, *p* = .011) and an approaching significant correlation between nose and eyebrows (*r*(22) = .40, *p* = .053). A positive correlation indicates that an increase in miss rates for one feature, was also observed for another. There were no other significant correlations between features (*p* > .16). As a Pearson's correlational analysis is only designed for linear relationships, a Spearman's rank correlational coefficient was also conducted, to check for non-linear relationships, which in this case revealed positive significant correlations between eyes and mouth, hair, outline (r_s(22) > .41, *p* < .047).

Target familiarity: To assess if a target's familiarity to participants affected miss rates, the overall mean miss rate (collapsed across feature) was compared to its familiarity to participants as indicated in the Spontaneous naming task in the Control section (familiar valid trials as a percentage of total trials shown). This would indicate how 'popular' the target face was to the participant pool. Results revealed an approaching significant correlation, (Pearson's *r*(22) = -.37, *p* = .08): the higher the likelihood of familiarity to participants, the lower the miss rate. A significant correlation would be expected here, as a higher familiarity target would be expected to yield a lower miss rate.

Explicit Choice: Overt recognition responses to the stimuli were collected in the form of an Explicit choice task (multiple choice, n=3) and compared to the Spontaneous naming task responses, for each item collapsed across feature condition. This was to establish if participants who were able to spontaneously name a target face were also likely to be able to overtly recognise the face. Explicit choice miss rates (%), for valid trials only, were compared to miss rates (%) for the Spontaneous naming task. As expected, there was a significant positive correlation (*r*(22) = .43, *p* = .035): the higher the spontaneous miss rate, the higher the explicit choice miss rate, as indicated by a scatterplot. This correlation is expected as it is likely that participants who could correctly recall the name of the target would subsequently choose the correct name in the Explicit choice task.

Lecturer version (1i)

The exact same methodology as that used for the celebrity version was repeated using lecturer target faces (see section 4.2 Method).

Participants: One hundred and forty eight participants (LJMU = 69, UCLan = 79) aged 18 or over (age: *M* = 21.1 yrs. [age not declared = 3] range = 18-38 yrs.) completed the online/lab based study (males n = 20, females n = 127, other n = 0, not declared n = 1).

Data: Data screening was the same as above (see section 4.3 Results) and collapsed across the two institutions. A check was made for missing data and fifteen stimulus images received no valid trials (across 5 items). This resulted in mean familiar (valid) trials of 3.5 per stimulus (total trials seen: M = 12.2). This meant that 28.9% of all trials were familiar and, therefore, valid for analysis (total valid trials = 316 trials (M = 52.7 trials per feature)). The low numbers of valid trials was expected as not all lecturers will have been familiar to the participants as they were recruited from different courses (and so not all students knew all the staff targets).

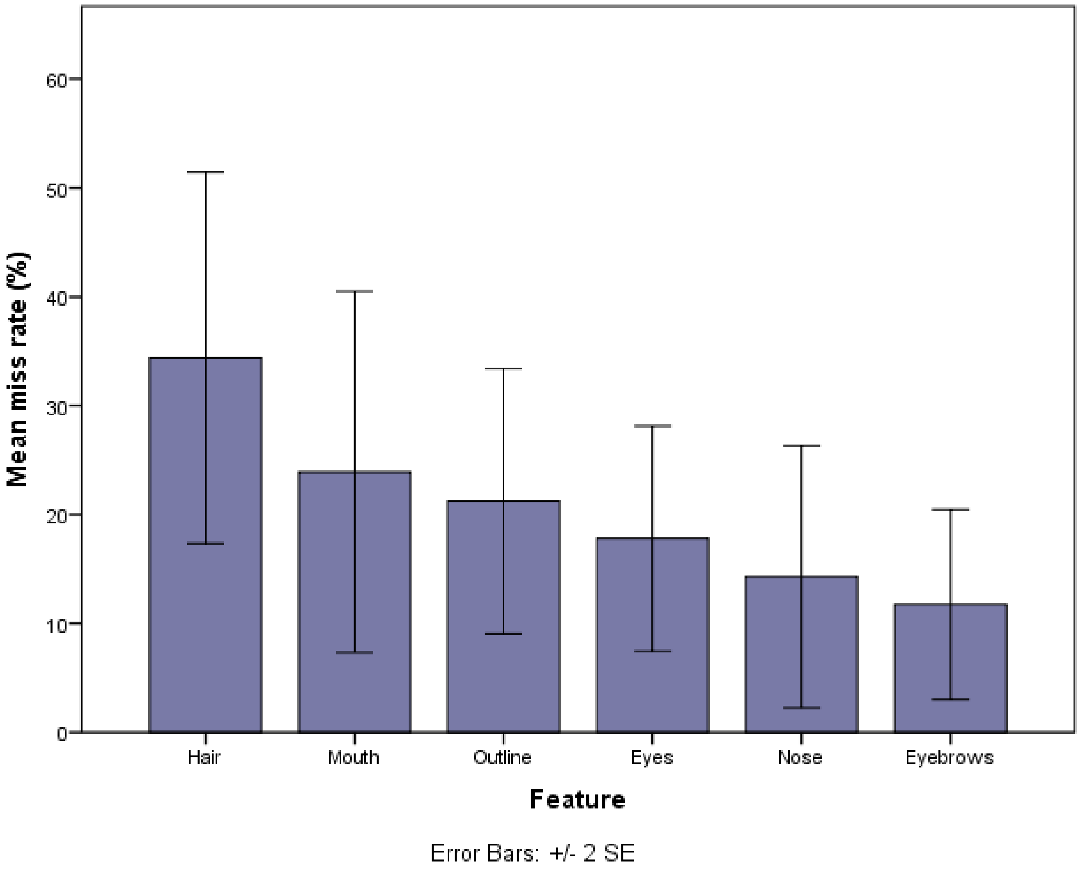


Figure 4.3-2: Phase 1i (lecturer) – Mean miss in percentage by feature.

Collapsed across 15 items. All mean values contained a minimum of 51 observations.

Descriptive analysis: Figure 4.3-2 shows that overall mean miss (%) scores were low across features (<34.5%, *M* = 21.1%, *SD* = 24.4%). The factor of feature yielded a miss rate hierarchy with hair (*M* = 34.4%, *SD* = 31.9%) by far the highest and eyebrows (*M* = 11.7%, *SD* = 15.1%) the least. Mouth was next highest (*M* = 23.9%, *SD* = 29.9%), similar to outline (*M* = 21.2%, *SD* = 22%) and eyes (*M* = 17.8%, *SD* = 17.9%). Nose was intermediate to eyes and eyebrows (*M* = 14.3%, *SD* = 19.9%). This pattern of effects is exactly the same as for the celebrity data, except that eyes yield the highest miss rate there, compared to the fourth highest here. In support of the feature saliency hierarchy, the number of items unaffected by the manipulation is, again, the inverse of the increase seen in the mean miss rates (%), except for eyes which show a lower number of unaffected items.

Analysis of Variance: Due to missing data (as there were no data available for analysis for fifteen cells, out of a maximum of ninety cells), feature was analysed as a between-subjects factor. An items analysis (female n=10, male n=5) with a between-subjects factor of feature (E,EB,N,M,H,O) and target sex (female, male) Univariate ANOVA showed no significant effect of feature (*F(*5,63) = 0.77 *p* = .57, η_p^2 = .06) or interaction with target sex (*F*(5,63) = 1.49, *p* = .21, η_p^2 = .11). Additionally, no main effect of target sex was found (*F*(1,63 = 0.42, *p* = .52, η_p^2 = .01). No stimulus characteristic ratings were collected for lecturer targets due to the limitations in recruiting participants familiar with the targets.

Simple contrasts were carried out against eyebrows (lowest miss rate) showing a significant difference between eyebrows and hair (*p* = .019) (highest miss rate), no other contrasts were significant (*p* > .20). Results for each feature were compared to each other showing a significant positive correlation between hair and nose (*r*(13) = .64, *p* = .033) and an approaching significant positive correlation between nose and outline (*r*(13) = .60, *p* = .068), different from the results of the celebrity version. There were no other significant correlations between features (*p* > .12). A Spearman's rank correlation coefficient revealed a significant positive non-linear correlation between hair and outline (r_s(13) = .41, *p* < .042).

Target familiarity: To assess a target's familiarity, the overall mean miss rate (collapsed across feature) was compared to its familiarity to participants as indicated in the Spontaneous naming task in the Control section (familiar valid trials as a percentage of total trials shown). Results showed no reliable correlation (*r*(15) = .14, *p* = .63). This result was not expected, as it was hypothesized that a higher level of familiarity would render targets more invariant to featural manipulations.

Explicit Choice: Explicit choice miss rates (%) for valid trials for the LJMU data could not be compared to Spontaneous naming miss rates as participants never incorrectly responded for the Explicit choice task in valid trials and so there was no variation in results. Upon extraction of only the UCLan data, only 4 items could be analysed in this way and were therefore too low to analyse.

Combined

For Phase 1i, celebrity (24 cases) and lecturer (15 cases) data were combined in a single analysis to give a broader and more general understanding of the effect of replacing target features.

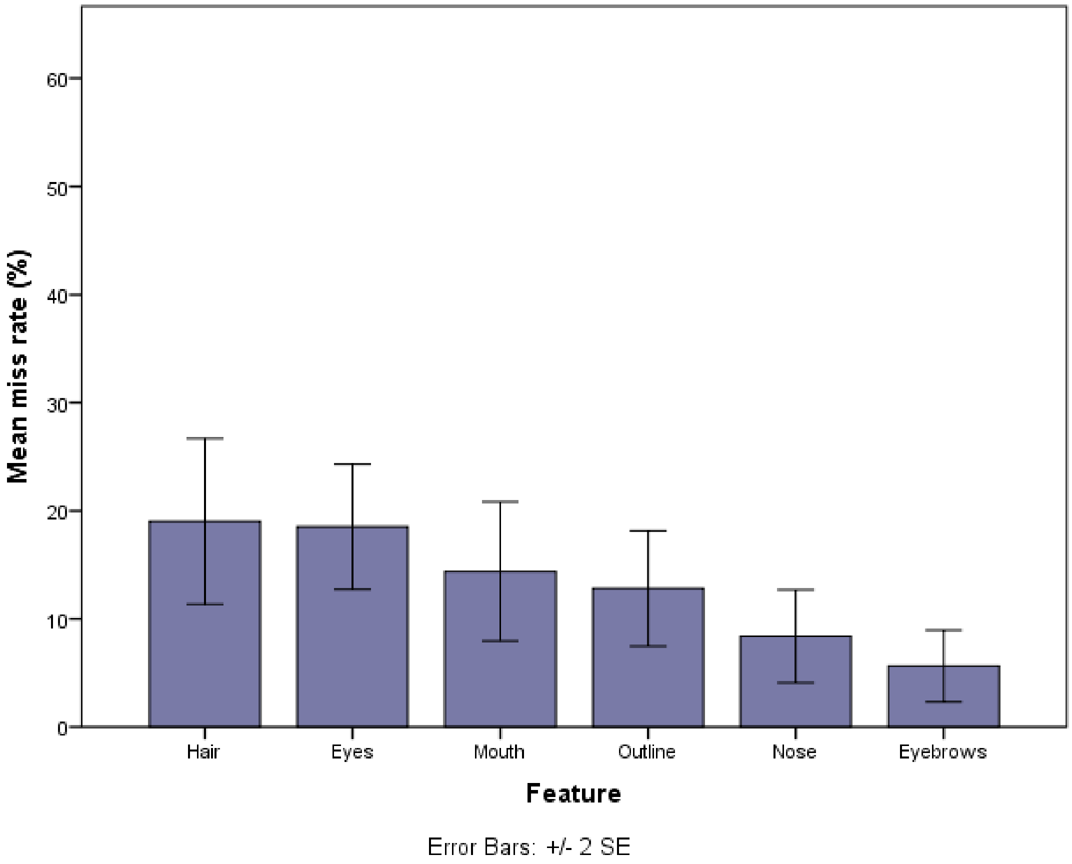


Figure 4.3-3: Phase 1i (Combined) – Mean miss in percentage by feature.

Collapsed across items (n=29)

Descriptive analysis: Figure 4.3-3 shows overall mean miss (%) rates collapsed across both versions of the experiment were low (*M* = 13.2%, *SD* = 17.8%). Overall, hair yielded the highest miss rates (*M* = 19.0%, *SD* = 23.7%). This was consistent with the individual lecturer data, but hair was the second highest in the celebrity version. This was a similar result to eyes which yielded the second highest miss rate overall (*M* = 18.5%, *SD* = 17.4%) which appears to be inconsistent with both the celebrity (1st) and the lecturer (4th) hierarchies (this was also the main inconsistency between the celebrity and lecturer data). Grouped in the middle was mouth (*M* = 14.4%, *SD* = 19.6%) and outline (*M* = 12.8%, *SD* = 16.2%). Nose (*M* = 8.4%, *SD* = 12.8%) and eyebrows (*M* = 5.6%, *SD* = 17.8%) yielded the lowest miss rates, consistent with both the celebrity and lecturer versions. In support of the feature saliency hierarchy, the number of items unaffected by the manipulation was the inverse of the increase seen in the mean miss rates (%), except that eyes yielded a lower number of unaffected items than the most salient feature, hair (H = 11, E = 8, M = 12, O = 16, N = 17, EB = 22).

Analysis of Variance: The Univariate ANOVA was repeated with features (6) as a between-subjects factor (39 items). Results showed a significant effect of feature ($F(5,213) = 3.51$, $p = .005$, $\eta_p^2 = .08$). Tukey's HSD post-hoc comparisons revealed significant differences between eyebrows (lowest miss rate) and hair (highest miss rate) ($p = .013$) and between eyebrows and eyes (2nd highest miss rate) ($p = .022$).

Results for each feature were compared to each other showing significant positive linear correlations (Pearson's) between outline and all other features ($p < .40$), except eyes. Additionally, significant positive correlations were found between eyebrows and nose and hair, nose and hair, and mouth and hair ($p < .050$). There were no other significant correlations between features ($p > .23$). Spearman's correlational analysis revealed significant positive non-linear correlations between eyebrows and nose and outline, between mouth and hair, and between hair and outline ($p < .047$). Approaching significant positive correlations were found between nose and hair and between mouth ($p = .083$) and outline ($p = .096$). There were no other significant correlations between features ($p > .15$).

4.4 Phase 1ii (Isolated version)

Introduction

Conducted only for celebrity targets, Phase 1ii showed individual target features presented in isolation to test if the feature saliency hierarchy results from Phase 1i would still be found. This also tested features outside of a whole face context, allowing for an observation on how features behaved without their surrounding inter-feature relationships. Celebrity and lecturer target faces had features sample from them and presented in isolation to form the composite stimulus set, which will be described in further detail below in the Method section (see 4.4 Method). The testing methodology for this stimulus set is outlined in the Common Methodology section (see 3.6 Common Methodology: Phase 1 and 2) where data was collected online in a face recognition task.

Method

Materials: The same six feature conditions were included as those in Phase 1i: eyes, eyebrows, nose, mouth, hair and facial outline. In this phase, the features were sampled from the target face and presented in isolation (no full-face context). Features were placed more centrally within the stimulus image (on the same blank grey background with the same dimensions as Phase 1i and 2) rather than their congruent replacement to avoid any remaining configural information that might have emerged in the spacing between the feature and the background so that only shape, and visual appearance (or texture) of features, was tested (see Figure 4.4-1).

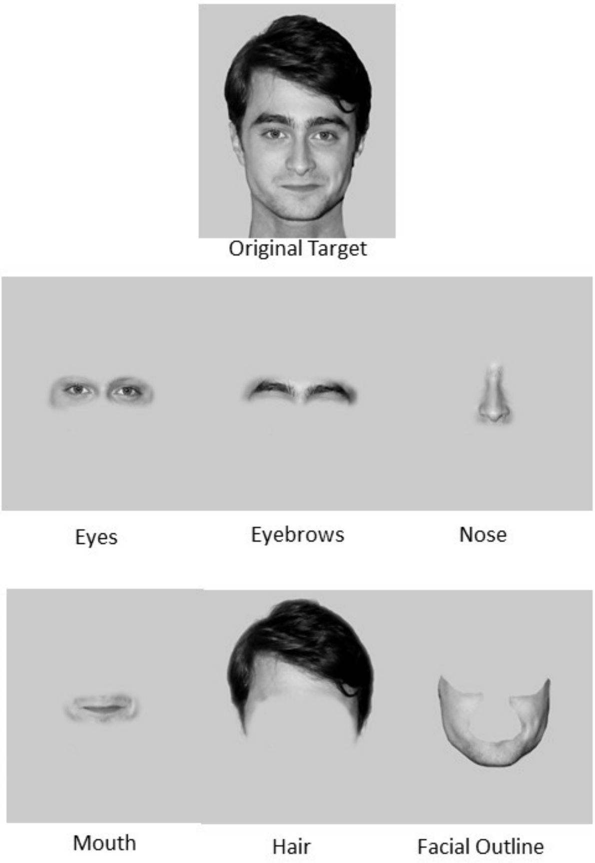


Figure 4.4-1: Phase 1ii – Example stimulus set

The six feature conditions generated by sampling the individual features from the target face (top) and presenting them in isolation (middle and bottom rows)

The exact same testing methodology as that used for the other experimental phases was repeated using isolated target features (see section 3.6 Common Methodology: Phase 1 and 2).

Participants: Eighty-four participants aged eighteen or over (age: *M* = 26.1 [age not declared = 1] range = 18-64 yrs.) completed the online study (males n = 24, females n = 59, other n = 1, not declared n = 0).

Results

Phase 1ii

Data: A check was made for missing data, of which no cases were found. This resulted in mean familiar (valid) trials of 12.0 per stimulus (total trials seen: *M* = 14.0). This meant that 85.9% of all trials were familiar and, therefore, appropriate for analysis (total valid trials = 1733 (*M* = 288.8 per feature)).

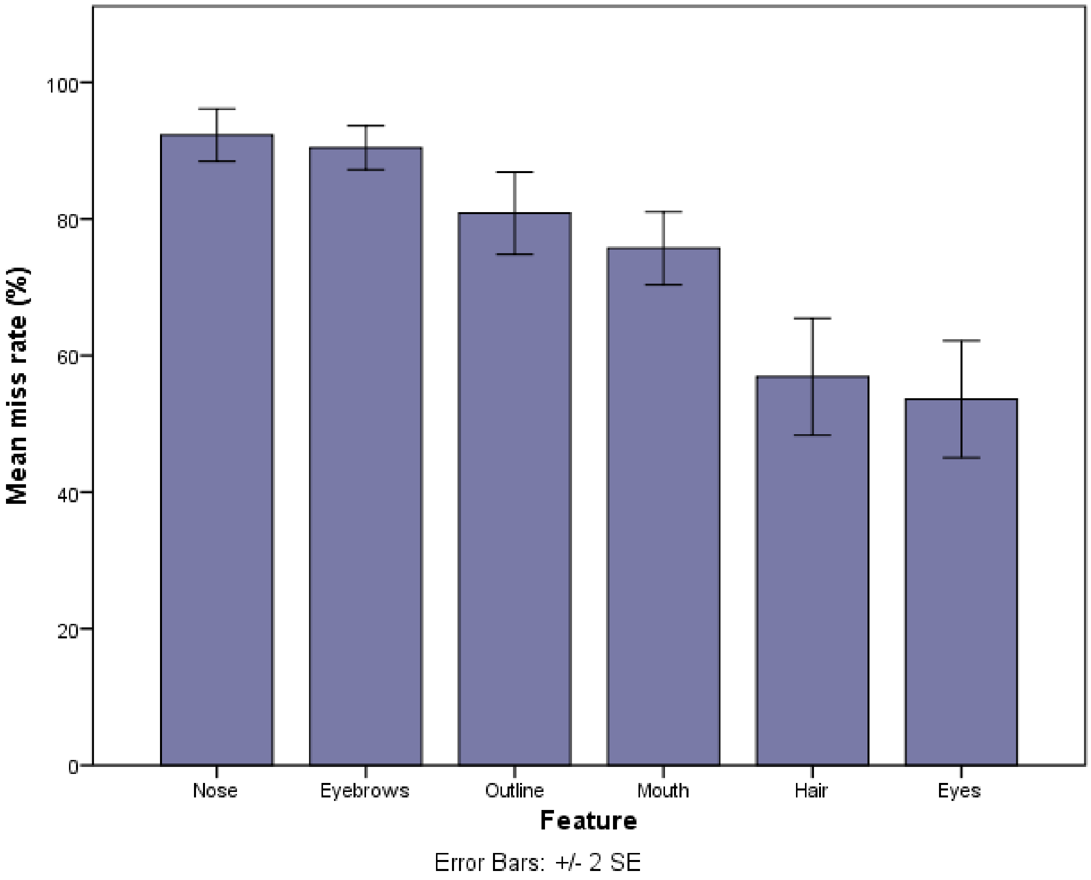


Figure 4.4-2: Phase 1ii (celebrity) – Mean miss in percentage by feature.

Collapsed across 24 items. All mean values contained a minimum of 275 observations.

Descriptive analysis: Figure 4.4-2 shows that overall mean miss rates across all conditions was high (*M* = 75.0%, *SD* = 21.3%) reflecting the much more difficult task. When shown in isolation, eyes yielded the lowest miss rates (*M* = 53.6%, *SD* = 20.9%), similar to hair (*M* = 56.9%, *SD* = 21.0%). Mouth was considerably higher (*M* = 75.7%, *SD* = 13.1%), similar to outline (*M* = 80.8%, *SD* = 14.8%). Eyebrows were higher (*M* = 90.4%, *SD* = 7.9%) similar to nose (*M* = 92.3%, *SD* = 9.4%), which yielded the highest miss rates near to ceiling level.

Analysis of Variance: For the full model, an items analysis (n = 24) (female = 12, male = 12) RM ANOVA for the within-subjects factor of feature (E,EB,N,M,H,O), between-subjects factor of target Sex (female, male) and covariate of memorability (scores taken from Ratings study) was carried out. This revealed a significant effect of memorability (*F*(1,21) = 6.04, *p* = .023, η_p^2 = .22). A scatterplot indicated a negative correlation between mean miss rate and memorability: the higher the memorability score, the lower the miss rates. Mean miss scores were collapsed across feature and correlated with the memorability covariate scores to analyse the overall significant effect of memorability. Results showed a significant negative linear correlation (Pearson's *r*(22) = -.53, *p* = .008). No significant interaction with feature was found (*F*(3.39,71.15) = 0.70, *p* = .57, η_p^2 = .03). Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated (X^2(14) = 29.0, *p* = .011) and therefore a Greenhouse-Geisser estimate was used (ε = .68).

The analysis was repeated separately for the covariates of attractiveness and trustworthiness. Results showed no significant effect of attractiveness (*F*(1,21) = 1.31, *p* = .27, η_p^2 = .06) or interaction with feature (*F*(3.41,71.51) = 0.74, *p* = .55, η_p^2 = .03). There was also no effect of trustworthiness (*F*(1,21) = 0.71, *p* = .41, η_p^2 = .03) or interaction with feature (*F*(3.50,73.54) = 0.39, *p* = .79, η_p^2 = .02).

To analyse the effect of feature and target sex, the covariate scores were removed from analysis in an attempt to increase statistical power. RM ANOVA revealed a significant main effect of feature (*F(*3.5, 76.97) = 28.4, *p* < .001, η_p^2 = .56, $X^2(14)$ = 27.8, *p* = .016, ε = .70). Post-hoc pairwise comparisons (Bonferroni corrected) were carried out to test for any significant differences between feature conditions. Eyes were significantly lower than all others (lowest mean miss rate) (*t*(23) > 7.63, *p* < .003), except hair. Nose (highest mean miss rate) was significantly higher than all others, except against eyebrows (*t*(23) > 3.07, *p* < .044). Mouth was significantly lower than eyebrows and nose and higher than the eyes and hair (*t*(23) > 4.81, *p* < .044). Again, hair was significantly lower than all features except eyes (*t*(23) > 4.30, *p* < .002). Results for each feature were compared to each other, showing a significant negative linear correlation between eyes and eyebrows (*r*(22) = .43, *p* = .037): the higher the miss rates for eyes, the lower for eyebrows. There were no other significant correlations between features (*p* > .11). A Spearman's rank correlation coefficient also revealed a significant non-linear negative correlation between eyes and eyebrows (r_s(22) > .46, *p* < .022) and an approaching significant negative correlation between hair and mouth (r_s(22) > .38, *p* < .071): the higher the miss rates for hair, the lower they are for the mouth. A main effect target sex was found to not be significant (*F(*1,22) = 2.98, *p* = .10, η_p^2 = .12).

Simple main effects: A significant interaction was found between feature and target sex (*F(*3.5, 76.97) = 2.65, *p* = .046, η_p^2 = .11). Simple main effects were calculated for each feature and target sex. Both males and females showed a significant effect of feature as expected (*p* < .001). For the effect of sex appearance on each feature, only hair was significant (*p* = .036) with females (*M* = 48.1%) yielding lower miss rates than males (*M* = 65.8%). This could account for some/if not all, of the significant interaction.

Target familiarity: To assess a target's familiarity, the overall mean miss rate (collapsed across replacement/feature) was compared to its familiarity to participants as indicated in the Spontaneous naming task in the Control section (familiar valid trials as a percentage of total trials shown). Results showed an approaching significant negative correlation (*r*(24) = -.36, *p* =.08): the higher the likelihood of familiarity to participants, the lower the miss rate. A significant correlation would be expected here.

Explicit Choice: Explicit choice miss rates (%) for valid trials only were compared to miss rates (%) for the Spontaneous naming task and there was no significant correlation (*r*(22) =.17, *p* = .43).

Phase 1i and 1ii

For the celebrity targets, Phase 1i and 1ii were able to be directly compared. Overall mean miss rates for Phase 1i (*M* = 9.1%, *SD* = 11.1%) were, as expected, considerably lower than for Phase 1ii (*M* = 75.0%, *SD* = 21.3%) which reflected the difference in task difficulty. Table 4.4-1 shows mean miss rates for features, collapsed across items for both Phases.

Table 4.4-1: Phase 1i and 1ii (celebrity) – Mean miss in percentage by feature

		Phase 1i		Phase 1ii	
		Mean	S.D.	Mean	S.D.
Feature	eyes	18.9	17.5	53.6	20.9
	hair	10.0	9.9	56.9	21.0
	mouth	9.2	7.3	75.7	13.1
	outline	8.3	10.0	80.8	14.8
	nose	5.7	6.6	92.3	9.4
	eyebrows	2.6	3.5	90.4	7.9
	Mean	9.1	11.1	75.0	21.3

A feature saliency hierarchy was observed in both phases with the results from Phase 1ii being the exact ordinal inverse of Phase 1i, except that eyebrows and nose are reversed.

An items analysis (female=12, male=12) within-subjects factor of feature (E,EB,N,M,H,O) and between-subjects phase of experiment (1i versus 1ii) RM ANOVA revealed a significant main effect of feature (*F(*3.5, 160.84) = 10.53, *p* < .001, η_p^2 = .19). Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated (X^2(14) = 57.74, *p* < .001) and therefore, a Greenhouse-Geisser estimate was used (ε = .70). A between-subjects effect of phase was also found to be significant (*F(*1,46) = 1660.54, *p* < .001, η_p^2 = .97) with miss rates higher for the isolated version.

Simple main effects: There was a significant interaction between feature and experiment phase (*F(*3.50, 160.84) = 33.18, *p* < .001, η_p^2 = .42). Independent samples t-tests was used to test whether any feature was affected differently by the experimental phase and if any experimental phase showed a different pattern of results by feature. Results showed that the interaction between feature and phase can be attributed to the varied increase in differences in miss rates between phases for features from the most to the least salient and due to the difference in task difficulty (p < .001), which will be discussed further later (see section 4.6 Phase 1 Discussion).

4.5 All Phase 1

To assess the overall effect of feature on recognition of the composites, all the data from Phase 1 was combined and analysed. This included the celebrity and lecturer data from Phase 1i as miss rates, and the data from the isolated Phase 1ii experiment (celebrity only) was inverted to Hits to become comparable with the data from Phase 1i. Due to missing data in the Lecturer version, feature was analysed as a between-subjects factor.

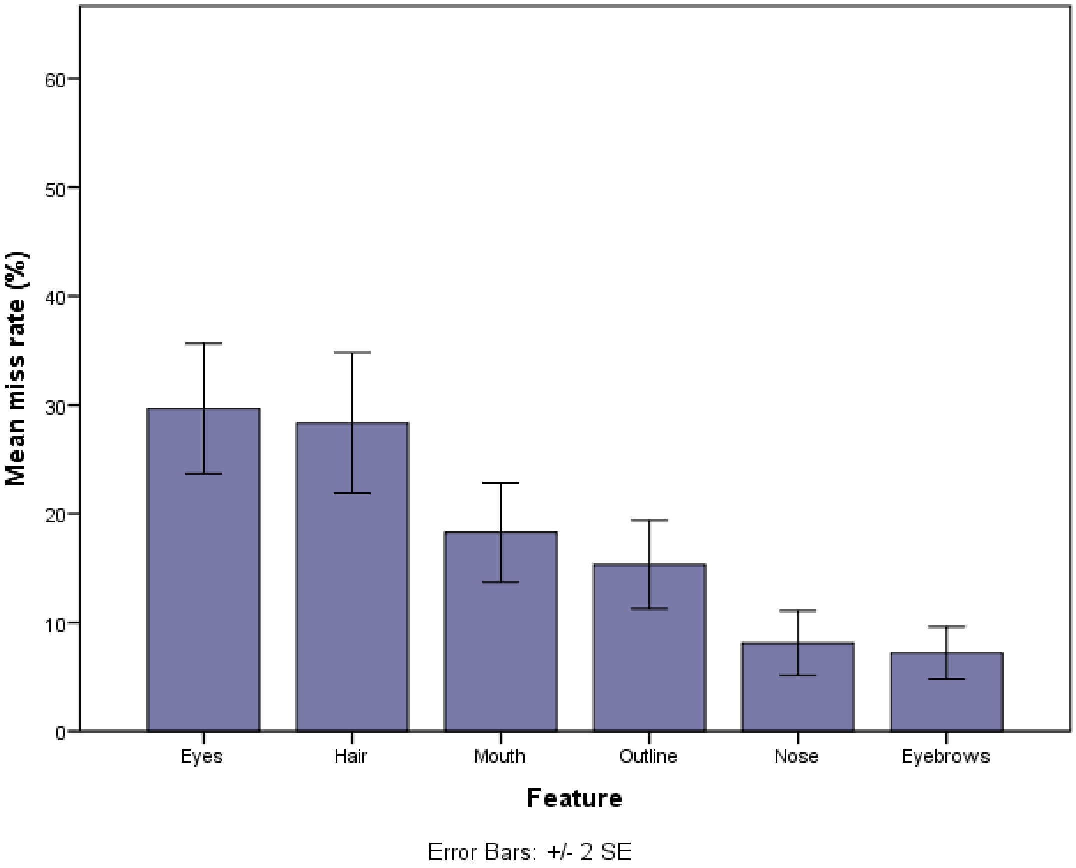


Figure 4.5-1: Phase 1 – Mean miss in percentage by feature.

Collapsed across minimum of 59 items. All mean values contain a minimum of 671 observations.

Descriptive analysis: Figure 4.5-1 shows overall mean miss (%) scores were moderately low (<29.8, M = 17.9 %, SD = 20.1 %). For the factor of feature, a clear order of miss rates was observed with eyes (E) (*M* = 29.7%, *SD* = 23.2%) yielding the highest and eyebrows (EB) (*M* = 7.2%, *SD* = 9.3%) the lowest overall. Hair (H) was next highest (*M* = 28.3%, *SD* = 23.2%), which was considerably higher than the mouth (M) (*M* = 18.3%, *SD* = 17.9%), and outline (O) (*M* = 15.3%, *SD* = 15.9%). Nose (N) (*M* = 8.1%, *SD* = 11.4%) was lower than outline and similar to eyebrows. Eyes' miss rate is no longer almost double the miss rate of the next highest feature as shown in the celebrity version of phase 1i.

Analysis of Variance: Univariate ANOVA was conducted with features (6) as a between-subjects factor (59-62 items). Results showed a significant effect of feature (*F*(5,357) = 16.91, *p* < .001, η_p^2 = .19). Tukey's HSD post-hoc comparisons revealed significant differences between eyebrows (lowest miss rate) and the top three highest miss rate features: hair, eyes and mouth (*p* < .013). Eyes (highest miss rate) were significantly higher than all other features (*p* < .009) except hair.

Results for each feature were compared to each other showing significant positive linear correlations (Pearson's) between hair and all other features (*p* < .027). Outline also replicated this result (p < .028) except for with eyes. Additionally, a significant positive correlation was found between eyebrows and nose (*p* < .018). An approaching significant positive correlation was found between eyebrows and mouth (*p* = .096). There were no other significant correlations between features (*p* > .20). Spearman's correlational analysis revealed significant positive non-linear correlations between hair and all other features (*p* < .007), except nose (*p* = .16). Mouth also followed this same pattern (*p* < .030) except with nose (*p* = .97). A further significant positive non-linear correlation was found between eyebrows and outline (*p* = .017). An approaching significant positive correlation was found between nose and eyebrows (*p* = .078). There were no other significant correlations between features (*p* > .10).

4.6 Phase 1 Discussion

Phase 1 was designed to test if there is a feature saliency hierarchy when target features from a familiar face were replaced with features from an unfamiliar face. This phase was designed as a pre-cursor to phase 2, so that any resulting feature hierarchy could be used to generate phase 2 stimuli. The intention was not to establish if identity could be concealed with the replacement of one feature, as this seemed unlikely, but rather to reveal if any differences in recognition occur by changing different features within a whole face.

The following themes emerged from the results:

1. Feature saliency hierarchy: Miss rates varied significantly between featural manipulations suggesting a feature saliency hierarchy.
2. Familiarity: Differences were found in the amplitude of miss rates for the different stimulus types. Familiarity with the target type (celebrity or lecturer) impacted on the results both quantitatively and qualitatively.
3. Whole-face context effect: As expected, the feature saliency hierarchy is inverted between the embedded whole-face version when target features are replaced, compared to when target features are presented in isolation.

Feature hierarchy results overall (all data) will be discussed first and the relative importance of features to recognition. Following sections will discuss the differences found between celebrity and lecturer data as well as levels of familiarity and a comparison of the whole-face embedded celebrity version of the experiment with the isolated version (phase 1ii). Overarching themes, such as potential limitations with the study and contribution to theory and practice, will be discussed in the General Discussion (see section 7 General Discussion).

Feature saliency hierarchy

Combined data (whole face versions): Feature saliency in the current study was interpreted from the mean miss rates, with a high miss rate interpreted as a large change to the identity of the target as a result of the feature being replaced. Thus, the higher the miss rate, the greater the saliency, and vice versa. Overall, mean miss rates across both versions of phase 1 combined (celebrity and lecturer) were low at 13.2% and demonstrate that replacing one facial feature does not impact greatly on recognition. The composite face effect (Young et al., 1987) effectively replaced half the face with another face, to change identity. A single feature replacement falls well short of this type of change, so it is unsurprising that miss rates were so low. However, results do show a pattern of miss rates so it seems that for some recognisers, replacing one feature is enough to change identity. This could be due to three factors:

1. Participants' familiarity with the target was robust enough in order to correctly recall the face, but perhaps not robust enough to be invariant to relatively small changes to the face (Clutterbuck and Johnston, 2005). This could be due to a short or unvaried familiarisation experience (Roark et al., 2003, Tong and Nakayama, 1999).
2. The participant may have been highly familiar with the target, with greater knowledge of the face, that meant greater holistic processing and a greater impact to recognition when changes were made to the whole face context (Tanaka and Farah, 1993).
3. Stimulus specific characteristics may have resulted in some stimuli being more susceptible to changes in identity through single feature replacement (Valentine, 1991, Valentine and Bruce, 1986).

The third possible reason was further explored by collecting ratings of target characteristics including 'memorability'. If a target face was memorable this may have made the face more or less vulnerable to changes made to the face depending on which feature(s) is making the face memorable and which feature is being changed (this will be discussed more in the section on Stimulus specific results). From the combined (overall) data, a feature saliency hierarchy was observed with hair replacement yielding the highest miss rates, followed by the eyes, and with the nose and eyebrows the lowest. In support of this result, the number of items (targets) not affected by the featural manipulation (0% miss rates) was almost the exact inverse of the feature saliency hierarchy found as indicated by miss rates (except the hair and eyes order changed). For example, the eyebrows were found to yield the lowest miss rates, and therefore yielded the highest number of items *not* affected by the manipulation.

The miss rate hierarchy indicates eyes as the second most salient which mostly aligns with previous literature that suggests eyes to be very, if not the most, salient for facial recognition (Logan et al., 2017, Schyns et al., 2002, Tanaka and Sengco, 1997). There are two possible reasons for this; the eyes are more salient because more time is spent encoding them for subsequent recognition due to a preference for eyes through communication and expression (Fraser et al., 1990, Haig, 1985). Or, the eyes physically provide more recognition cues due to their combination of both high and low frequency information, high level of contrast, use for communication and expression (Peterson et al., 2007, Schyns et al., 2002). In contrast, hair is normally associated with an increased saliency in unfamiliar, not familiar face recognition (Young et al., 1985), but the results here suggest otherwise. It is possible that the differing types of familiar stimuli may have impacted on this hierarchy due to the ways in which the faces were familiarised (this will be discussed further in section4.6 Familiarity). The distribution of features across the saliency spectrum was supported by post-hoc results showing significant differences between these two most salient features and the eyebrows, the least saliency feature. No significant differences were found between other features, suggesting the distribution of saliency was not overly spread out and this may in part be due to the overall low miss rates not allowing for large differences between features and the relatively large standard deviation scores. The mouth was found to be the third most salient which mostly aligns with previous research (Davies et al., 1977, Haig, 1986, Fraser et al., 1990) and may, in part, be due to a heightened focus towards the mouth during communication.

The current study found eyebrows to have little effect on recognition when the shape was changed, thus maintaining contrast patterns (intensity and thickness). This supports the study by White (2004) where plasters were used to obscure the eyebrows but leave a small difference in contrast between the eyebrows and surrounding skin. They found recognition rates did not decrease as much as when the eyebrows were filled in with surrounding skin (contrast removed). Therefore, it can be assumed that eyebrows are extremely salient in terms of their relative role in the contrast pattern of the face but not for featural information.

Positive linear correlations were found between the outline and all other features, except the eyes, which suggests that the outline aligns with increases/decreases in miss rates with all other features. For example, if the mouth miss rate is high in target 1 and low in target 2, then this pattern will also be true of the outline. This result seems sensible as the outline provides the context for all other features and is perhaps the container for all configural information as it fills the spaces between most of the other features (Tanaka and Farah, 1993). It also borders all of the features and these boundaries, which have been selected by the experimenter, may be areas of interaction and interrelationships between features, facilitating an influence of each feature upon the facial outline. Other correlations included the hair with four other features: the nose, eyebrows, outline and mouth. Again, like the outline, the hair may provide some context for other features.

All phase 1 data: The analysis across all of the phase 1 experiments found a very slightly different feature hierarchy to that of the combined (celebrity and lecturer) results discussed above. This further analysis also included the inverted results from the isolated version to give a broader understanding of the overall saliency of each feature for recognition. Unsurprisingly, eyes were the most salient of all features, followed by the hair, with the eyebrows the least. Therefore, a difference occurs in the saliency of the eyes and hair between the analyses, with the addition of isolated data adding more weight to the importance of the eyes. This perhaps reflects the amount of detail available from the eyes (in comparison to the hair) which may become more important when presented in isolation. It may also suggest that eyes are more independent of their interrelationships with the other facial features.

If the formal aspects of each feature are taken into consideration, in terms of the amount of information available, then this may reflect differences in the importance of features (Peterson et al., 2007). This may include shape from shading information, contrast across the feature and whether it’s dynamic. The relative levels of these attributes for feature could be reflected in the saliency hierarchy results (Peterson et al., 2008). Features that contain more varied information, such as shape from shading (indicated by volume), high contrast levels, and dynamic streams, could be found to be more salient: in the current study, this included the hair, eyes, and mouth. Featural manipulations in the current study may only have affected some shape from shading of the feature, rendering the other types of information mostly unchanged.

Familiarity

Differences were found in the amplitude and hierarchy results between the celebrity and lecturer versions suggesting that the type of familiarity has an effect on miss rates, both quantitatively and qualitatively.

Performance: Overall mean miss rates were low for both celebrity (9.1%) and lecturer (21.1%) versions of the experiment, both supporting the notion that changing a single feature in a familiar target face is not sufficient conceal identity all of the time. However the lecturer mean was considerably higher than that of the celebrity (*MD*=12.0%). One possible explanation for this is a difference in level of familiarity. The experiments were designed with a familiarity threshold of correctly recalling the name of the identity. Only trials that met this threshold were used for analysis. However, it is possible that this threshold actually represents a fairly low level of familiarity. It is likely that the celebrity targets had been familiarised over a longer period with various difference guises and appearances in comparison to the lecturer data, rendering them more invariant to the featural manipulations. In support of this, the overall effect of feature was found to be significant for the celebrity data but not for the lecturers. The overall very low miss rates for lecturer data imply an effect of manipulating parts of the face, but it seems that this effect is not vastly different between which feature is being manipulated and as such suggests a lower level of familiarity.

Feature hierarchies: Although there was a differences in the effect of feature manipulation between the different types of target faces, there were many similarities between the feature saliency hierarchies. The celebrity and lecturer data followed the same hierarchy pattern, except for one main difference: the eyes, which were the most salient for the celebrity data (and significantly different from all other features, except the hair), but fourth for lecturers. It is possible that, in addition to the miss rate amplitude differences between the types of target faces, there appears to be some qualitative differences that have resulted in differences in saliency for the eyes. This could be explained by the method of familiarisation whereby lecturer faces may well have been familiarised from a large viewing distance (e.g. lecture theatre), and, therefore, preferences for encoding may have been given to the more obvious features visible from those large distances, hair and outline, rather than the less visible eyes, and that eyes occupy little space on the face in comparison to hair. It is also possible that the lower levels of familiarity have replicated findings in the literature that suggest a preference for the external features for unfamiliar or newly familiar faces (Clutterbuck and Johnston, 2007, Clutterbuck and Johnston, 2002, Ellis et al., 1979, Young et al., 1985). The results of this study clearly reflect an external features preference (1st hair; 3rd outline). A lot of teaching, especially during lectures, takes the form of the lecturer speaking to the audience where the mouth is inevitably highly animated during this process, and visible from a long distance, resulting in familiarisation for this feature, which is reflected in the high saliency scores for this feature (2nd). The mouth also animates the external contours of the face and may provide an interaction with the contour that has supported the outlines importance. However, mouth is still the third most salient for the celebrity data, and may have only shifted to 2nd for lecturers because of the decrease in miss rates for eyes. The range of responses was also larger for the lecturer than celebrity data. Two features yielded a full range of miss rate between 0.0% and 100.0%, suggesting a slightly more binary response signal due to the manipulation having a different effect depending on the target. It is also possible that there were vast differences in familiarity levels across targets in the lecturer versions that have allowed for a large range of miss rates between targets.

The two types of targets revealed positive correlations between different features suggesting that any encoding links between features were modulated by the degree of familiarity. Celebrity data revealed a positive linear correlation between the eyes and outline suggesting that they followed the same pattern of manipulation effects for targets. They are relatively separate on the face in terms of location except for the superior border of the outline, lies beside the inferior border of the eyes. Therefore, any deformations created by either feature, would potentially impact on the other, influencing the overall perception of the face through context effects (Tanaka and Farah, 1993). In contrast, the lecturer data showed a positive linear correlation between the hair and nose, which are mostly unrelated on the face.. A positive non-linear correlation was also found between the hair and outline. These two features could both be considered external features and are also located together and form the whole of the face/head border. A further non-linear correlation was found between the eyes and mouth for the celebrity data, perhaps influenced by their combined use in expression and communication.

Trials: For item familiarity, the number of valid trials for the celebrity version was quite high at 87.2% of the total trials for Phase 1i and 85.9% for Phase 1ii, suggesting that the celebrity target selection was suitable to the participant pool. In contrast, the lecturer version of the experiment yielded valid trials of only 28.9% of the total trials shown. There are two apparent reasons for this:

1. Within the LJMU experiment, targets were recruited from two different departments but included in the same study. Therefore, it is likely that students from different departments will have not been familiar with the targets from the other department.
2. Even though participants were recruited from student pools where the lecturers taught, it is possible that students may not have engaged with those specific lecturers, reducing the numbers of valid trials. It was apparent that this was occurring through early preliminary assessment of the data collection process. When the proposed sample size of ninety participants was reached, the analysis showed a very low level of valid trials. Therefore, recruitment continued above the estimated sample size, so that 148 participants were recruited.

Further analysis was carried out to assess whether the likelihood of familiarity for each target had an effect on miss rates. Mean miss rates were compared to the number of valid trials per target. The latter may suggest some level of ‘popularity’ with the participant pool as an indication of familiarity and to analyse if this level of familiarity affected miss rate results. All phase 1 experiments showed no significant correlation between miss rates and the level of familiarity per target. However, both phase 1 (celebrity) experiments (isolated and embedded) revealed an approaching significant correlation, which would be more expected. The lack of correlation in the lecturer data may suggest that a target’s popularity (or level of familiarity) does not affect miss rates under the stimulus manipulations. There are clear quantitative and some qualitative differences in miss rates between the different versions of the experiment and it is possible that this may be driven, in part, by the level of target familiarity, indicated through the number of valid trials and the popularity of the targets being tested.

Whole face context

Phase 1 included manipulated features within a whole-face context to test if a saliency hierarchy existed. Further to this, an isolated feature version (phase 1ii) was carried out, only using celebrity targets, as a baseline from which to compare the effect of a whole-face context. The overall effect of feature was found to be statistically significant for both, reinforcing a difference in miss rates between different target features being presented. Overall mean miss rates for the isolated phase were high at 75.0%which is a 65.9% increase in miss rates from the embedded phase (phase 1i, *M*=9.1%) and reflects the difference in task difficulty, supported by a significant effect of experimental phase in the ANOVA. Although comparable in the features manipulated between the phases, the embedded phase replaced the target feature, whereas the isolated phase presented the target feature in isolation. This suggests that replacing a target feature in a whole face is less detrimental to recognition, as indicated by the low miss rates in phase 1i, compared to showing those replaced features in isolation and out of a whole face context. For the isolated feature phase, an almost exact inverse of the feature saliency hierarchy results from the embedded phase was found. The only exception to this order was that the nose and eyebrows were switched in hierarchy replacement, although still at the bottom of the saliency hierarchy and with very small differences in miss rates.

Overall standard deviation for the isolated version was moderately low at 21.3%. This is probably as a result of data approaching a ceiling effect. However, the feature SD scores mostly followed the inverse of the pattern of feature saliency for miss rates. The features with the lowest miss rates (eyes and hair) showed the largest standard deviation (>20.8%). The eyes showed a large range in miss rates (0-92.9%) suggesting a high range of variability across items. The highest scoring features yielded the smallest standard deviation range (nose=9.4%; eyebrows=7.9%). This suggests that the nose and eyebrows are not only the least salient for recognition in terms of their high miss rates, but they are more consistently found to yield this result across all items. This pattern was not found for the whole-face phase where the standard deviation scores varied in whether they were higher or lower than the mean and did not follow the pattern of the saliency hierarchy.

The overall high miss rates for the isolated version have allowed for large differences between the feature conditions. Post-hoc pairwise comparisons revealed the eyes and hair to be significantly different from all other features, except each other, thus grouping them at the top of the saliency hierarchy. The nose, with the highest miss rates, was significantly different from all other features except the eyebrows, again grouping those two features at the bottom of the saliency hierarchy. These results are similar to the pairwise comparisons in the embedded version that grouped the two highest scoring features (upper face) and the two lowest (lower). The isolated phase also showed a significant difference between the mouth and all other features, except the outline. This suggests that the middle group of features for the isolated version were more spread out than those found in the embedded version, and are statistically further away from the two most salient features (eyes and hair).

For the isolated phase, correlational analysis revealed signification linear and non-linear negative correlations between the eyes and eyebrows suggesting that the two features follow an opposite pattern of miss rates across the different target faces (items); when eyes are more salient, eyebrows will be less salient for the target face. This relationship could be considered surprising given their close physical location on the face and the movement of one feature often affects the other (Sadr et al., 2003). In contrast, it is not surprising given that eyes are considered the most important for recognition. It is possible that eyebrows, due to their physical proximity, may change the appearance of the eyes to some extent, but this is secondary to changing the eyes themselves. This pattern of results was not found for the embedded phase of the experiment, suggesting that the relationship between the two features only affects miss rates for features presented in isolation.

Summary

Overall mean miss rates were low with a feature saliency hierarchy detected that showed the hair and eyes to be the most salient, with the nose and eyebrows the least. The celebrity and lecturer versions of phase 1i results yielded similar feature saliency hierarchies except for one feature: the eyes (celebrity = 1st; lecturer = 4th). This could be explained by differences in familiarity levels or focus of attention at encoding. The nose and eyebrows were consistently found to be the least salient. Phase 1ii (isolated features) showed an almost exact inverse of the results from the counterpart embedded version (celebrity Phase 1i) as expected with the higher miss rates detected in the isolated version reflecting the difficulty in recognising features outside of a whole-face context.

A discussion of the effects of stimulus specific characteristics, explicit choice and the limitations of the design and methodology are outlined in section 7. General Discussion.

5 Phase 2

5.1 Introduction

Phase 2 presented participants with the target face incrementally replaced with unknown features from the counterpart unknown unique composite face (compound effect) so that each of the six stimulus conditions represented a replacement of replacement (1-6). More details of this can be found in the method section below (see section 5.2 Method). Phase 2 was repeated for both celebrity and lecturer targets. The testing methodology for this stimulus set is outlined in the Common Methodology section (see 3.6 Common Methodology: Phase 1 and 2) where data was collected online in a face recognition task.

5.2 Method

Materials: The feature saliency hierarchy results from Phase 1i (celebrity version) were used to generate the phase 2 stimuli. Due to the chronology of data collection, the hierarchy results from only the celebrity version of the experiment were used to generate the stimuli for phase 2. There was one difference between the lecturer hierarchy and the celebrity version (eyes were the 4th most salient feature for lecturers but shifted to 1st in the celebrity data and in turn swapped replacements with the hair). However, the rotational starting point of feature replacement should have mitigated any effect of the order change between eyes and hair.

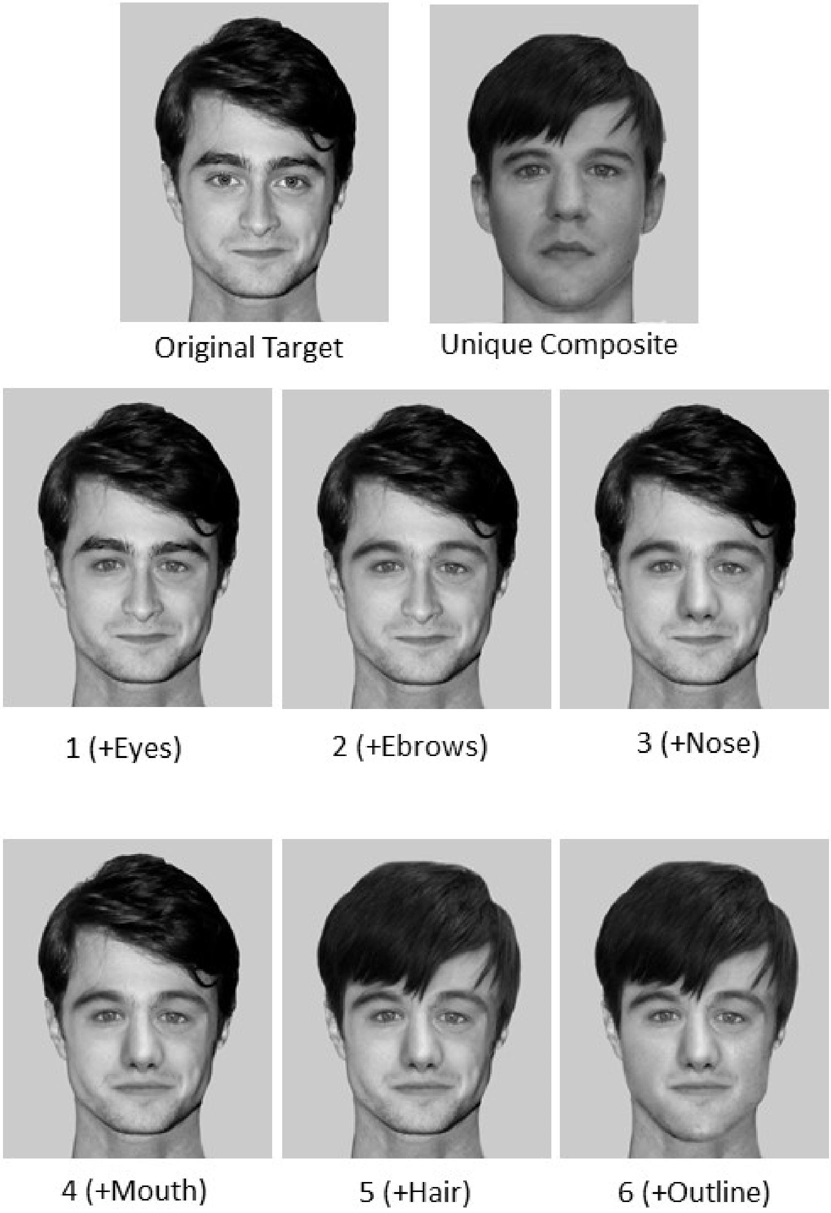
The feature saliency hierarchy was as follows (from lowest miss rates to highest miss rates): eyebrows, nose, outline, mouth, hair and eyes. This order was used in the replacement of features with the starting point staggered, resulting in six different configurations (see Table 5.2-1).

Table 5.2-1: Rotational configurations of feature replacement for phase 2

Position	Rotation					
	A	B	C	D	E	F
1	Eyebrows	Nose	Outline	Mouth	Hair	Eyes
2	Nose	Outline	Mouth	Hair	Eyes	Eyebrows
3	Outline	Mouth	Hair	Eyes	Eyebrows	Nose
4	Mouth	Hair	Eyes	Eyebrows	Nose	Outline
5	Hair	Eyes	Eyebrows	Nose	Outline	Mouth
6	Eyes	Eyebrows	Nose	Outline	Mouth	Hair

In order to prevent any cut-off points that might occur (for example, a large decrease in recognition may occur after replacing feature 4) and to mask any changes after that point, a rotation of order starting points was adopted to give different configurations as part of a Latin Square design. For example, one target face was replaced with the eyes, then the eyebrows and nose etc. Another target face was replaced with the eyebrows, nose and outline etc. If the starting order stayed the same for every target, then a criterion point might have been found from old to new face mid-way through the feature replacement, and therefore making any further replacement of features redundant. Systematically rotating the starting point in the feature replacement order meant all features had a chance of being replaced before and after any criterion point that may have been found.

Each target was assigned one configuration order. For the celebrity version of the experiment (n = 24), this resulted in each configuration being assigned to four targets (2 females, 2 males). For the lecturer version of the experiment (n = 15), uneven target samples from two institutions resulted in uneven assignment of the six configurations, but with the aim to evenly balance between the two institutions and between sex appearance of the targets (LJMU: A = 1(1 female), B = 2 (1 female, 1 male), C = 2 (1 female, 1 male), D = 2 (1 female, 1 male), E = 1(1 female), and F = 1 (1 male). UCLan: A = 1 (female), B = 1 (male), C = 1 (female), D = 1 (female), E = 1 (female), and F = 1 (female)). Due to the rotating starting order of feature replacement, the replacement stimulus will have differed in which features had been replaced at that point, across the whole stimulus set (see Table 5.2-1 for the rotational order of features by replacement). However, selection of the stimulus to be shown in the experiment, was only based on the stimulus' replacement, randomly selected for each participant. An example stimulus set for a target face can be seen in Figure 5.2-1.



Original Target

Unique Composite

1 (+Eyes)

2 (+Ebrows)

3 (+Nose)

4 (+Mouth)

5 (+Hair)

6 (+Outline)

Figure 5.2-1: Phase 2 – Example stimulus set

Six replacement conditions (middle and bottom rows) created by gradually replacing the target face (top left) with features from the unique composite (top right) (showing Configuration F). Feature replacement order was dictated by the feature saliency hierachy results from Phase 1i.

5.3 Results

All data was screened and scored in the same way as for Phase 1 (see section 4.3 Results for more detail).

Celebrity version

Participants: Ninety-one participants aged 18 or over (age: *M* = 31.0 yrs. [age not declared = 4] range = 18-60 yrs.) completed the online study (males = 31, females = 59, other = 0, not declared = 1).

Data: A check was made for missing data, of which no cases were found. This resulted in mean familiar (valid) trials of 12.9 per stimulus (total trials seen: *M* = 15.2). This meant that 85.1% of all trials were familiar and, therefore, appropriate for analysis (total valid trials = 1859 (*M* = 309.8 per replacement)).

Table 5.3-1: Phase 2 (celebrity) – Mean miss in percentage by feature and replacement

		feature						
		eyes	eyebrows	nose	mouth	hair	outline	Mean
replacement	1	5.9	3.5	3.3	11.9	18.9	6.6	8.3
	2	27.2	11.8	7.6	17.7	16.9	21.7	17.1
	3	54.2	44.5	14.6	26.6	34.9	22.5	32.9
	4	67.3	68.7	73.8	24.5	62.0	28.1	54.1
	5	79.8	62.3	77.7	48.7	51.7	82.7	67.2
	6	81.5	87.5	69.5	95.0	90.9	92.1	86.1
	Mean	52.0	47.3	43.3	40.5	46.9	44.2	45.7

Mean miss rates (%) of each feature at their replacement point, collapsed across items (4 items per mean of feature and replacement). All mean values contained a minimum of 42 observations.

Descriptive analysis: Table 5.3-1 shows a strict ordinal increase for the overall mean of replacement (see last column). Eyes and outline follow this ordinal pattern; the remaining features do not, as can be seen at replacement 2 for hair, replacement 4 for mouth, replacement 5 for eyebrows and replacement 6 for nose. Miss rates were generally quite low at replacement 1 and quite high at replacement 6. Overall mean miss rates (last row) were somewhat similar across features (range = 40.5-52.0%) with the highest miss rates for eyes, followed by eyebrows, hair, outline, nose and mouth. For replacement 6, it was hypothesised that replacing all six features would result in 100% miss rates. However, the total (not overall Mean) miss rate for replacement 6 was 86.6%. The remaining responses showed that 9.3% were recorded as 'familiar' and 3.8% were correct recognitions (n = 11 responses across four items).

Analysis of Variance: Due to only using valid trials which was dictated by 100% correct recognition for veridical images, it can be assumed that a veridical image will yield 0% miss rates and that a different image would produce 100% miss rates. To be able to compare replacing one feature to veridical images and replacing all features (replacement 6) to a different image, a baseline score of zero and ceiling score of 100 was used for items, giving a total of eight levels for replacement. For the full model, a Univariate ANOVA for the factors of feature (being replaced at that point) (E,EB,N,M,H,O) x replacement (0-7) x target sex (female, male) with a covariate of the memorability scores per target (items) was carried out. This was to test for the overall effect of facial characteristic ratings on miss rates. Results showed no significant main effect of target sex (*F*(1,95) = 1.61, *p* = .21, η_p^2 = .02) or covariate of target memorability (*F*(1,95) = 0.10 *p* = .75, η_p^2 = .00). The analysis was repeated separately for the covariates of attractiveness and trustworthiness for target ratings revealing no effect of attractiveness (*F*(1,95) = 0.70, *p* = .41, η_p^2 = .01) but a significant effect of trustworthiness (*F*(1,95) = 8.41, *p* = .005, η_p^2 = .08) [2], and a scatterplot indicated a negative correlation between mean miss rate and trustworthiness: the higher the trustworthy score, the lower the miss rates. To analyse the overall significant effect of trustworthiness, mean miss scores were collapsed across feature to give an overall miss rate for each target. Using a correlations analysis, these scores were then analysed against the trustworthiness covariate scores to see how trustworthiness of a target face affected its miss rates. Results showed no significant correlation (Pearson's *r*(22) = -.12, *p* = .57).

To increase power, the factor of target sex and covariates of target characteristic ratings were removed from analysis and the test repeated to look for any effects of which

[2] Including covariates reduces the power of the test. Therefore, the analysis was also rerun, without the addition of positions 0 and 7 and the results were the same: not significant for memorability and attractiveness, but significant for trustworthiness.

feature had been replaced and how miss rates were affected by the amount of features having been replaced. The Univariate ANOVA revealed a significant main effect of replacement (*F(*7,144) = 88.16, *p* < .001, η_p^2 = .81) but not feature (*F(*5,144) = 1.34, *p* = .25, η_p^2 = .04). An approaching significant interaction was found between feature and replacement (*F(*35,144) = 1.48, *p* = .058, η_p^2 = .26).

Tukey HSD post-hoc tests were used to assess the pattern of replacement miss rates in terms of their distribution and as part of that test, Tukey groups levels together that are not significantly different from one another as part of a homogenous subset output table. If replacement levels are significantly different from one another, they are placed in a separate subset.Table 5.3-2 shows replacements mostly paired together.

Table 5.3-2: Phase 2 (celebrity) - Homogenous subsets for replacement

replaceme		Subset			
nt	N	1	2	3	4
0	24	0.0			
1	24	8.3			
2	24	17.1	17.1		
3	24		32.9		
4	24			54.1	
5	24			67.2	
6	24				86.1
7	24				100.0
Sig.		.05	.10	.27	.20

Tukey HSD, means for groups in homogeneous subsets are displayed. Based on observed means. Alpha = .05. Replacement "0" represents no change (0.0% miss) and 7 for 100.0% miss. Results show differences between feature replacement in a paired stepwise manner (Between-subjects factors of feature and replacement).

As can be seen, there is a very consistent result: at least three features need to be replaced before a significant difference occurs in miss rates between the next feature replacement. To assess for the size of the effects between the subsets, Cohen's d was calculated from one subset to the next and found all three differences (Subset 1-2, 2-3, 3-4) showed a large effect size (*d* > 0.91).

As replacement was significant in the Univariate analysis and was the most effective factor in the experiment, data were then grouped by replacement to allow feature to become a within-subjects factor, increasing the power of the analysis. RM ANOVA for the within-subjects factor of feature (6) and the between-subjects factor of replacement (8) showed no significant effect of feature (*F*(5,120) = 1.49, *p* = .20, η_p^2 = .06). Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated (χ^2(14) = 0.49, *p* = .33). Replacement was, as before, significant (*F*(7,24) = 57.94, *p* < .001, η_p^2 = .94).

Table 5.3-3: Phase 2 (celebrity) - Homogenous subsets of replacement (Repeated measures)

replacement	N	Subset					
		1	2	3	4	5	6
0	4	0.0					
1	4	8.3					
2	4	17.1	17.1				
3	4		32.9	32.9			
4	4			54.1	54.1		
5	4				67.2	67.2	
6	4					86.1	86.1
7	4						100.0
Sig.		.25	.34	.08	.56	.16	.49

Tukey HSD, means for groups in homogeneous subsets are displayed. Based on observed means, and Alpha = .05. Results show a stepwise pairing of features. Three features need to be changed before a significant drop in recognition occurs (Within-subject factor of feature, between-subjects factor of replacement).

Homogenous subsets shown in Table 5.3-3 indicate that three features need to be replaced before miss rates start to significantly rise, and further replacement replacements require only two changes for a significant increase. One noticeable difference can be observed between the overall homogenous subset for the between-subjects (feature and replacement) analysis (Table 5.3-2) compared to the RM subsets (Table 5.3-3): The between-subjects analysis shows the data grouped into 4 subsets, but the data is spread out more into 6 subsets for the RM analysis. Cohen’s d found a medium effect size for differences between subsets 2-3 (*d* = 0.67), 3-4 (*d* = 0.56) and 4-5 (*d* = 0.57) and a large effect size for differences between subsets 0-2 (*d* = 0.94) and 5-6 (*d* = 0.75). The design aimed to detect a medium effect size, should one exist, and this analysis illustrates that this aim was supported. In the more powerful analysis here, there are more subsets with lower effect sizes needed for a significant difference to occur.

Cohen's d was also calculated for the effect sizes between replacement positions and yielded large effect sizes between all replacements (*d* > 0.70) except a medium effect size between replacements 5 and 6 (*d* = 0.44) and a medium to large effect size between 6 and 7 (*d* = 0.63) (overall mean, *d* = 0.68).

Simple main effects: A significant interaction was found between feature and replacement ($F(35,120) = 1.65$, $p = .024$, $\eta_p^2 = .33$). Simple main effects were used to establish if any feature was affected differently by its replacement number, and if any replacement was affected differently by which feature was being replaced at that point. Results revealed that features showed a significant effect of replacement, as expected ($p < .001$). For the effect of which feature was being replaced at the replacement number, only replacement 4 was significant ($p = .044$). At replacement 4, mouth yielded the lowest miss rates and was therefore used for Simple contrasts against the other levels (features) to test for any significant differences between features. Mouth was significantly lower than nose ($F(1,3) = 38.28$, $p = .009$, Bonferroni corrected, α_{altered}=.05/4 = .0125). The configurations where mouth (24.5%) and nose (73.8%) were replaced at replacement 4 show one main difference: mouth follows the replacement of less impactful features (outline, nose and eyebrows), whereas the highest miss rate for nose at replacement 4 has already had potentially more impactful features replaced (eyebrows, eyes and hair). The interpretation of this potentially important result will be discussed further in the Phase 2 Discussion (see 5.4 Phase 2 Discussion). Post-hoc Tukey HSD Homogenous subsets were calculated for each feature. Given the value of these subsets for interpretation of the results, all six subsets are presented in Table 5.3-4.

Table 5.3-4: Phase 2 (celebrity) – Homogenous subsets of replacement by individual feature

(a) eyes

		Subset			
Pos.	N	1	2	3	4
0	4	0.0			
1	4	5.9			
2	4	27.2	27.2		
3	4		54.2	54.2	
4	4			67.3	67.3
5	4			79.8	79.8
6	4			81.5	81.5
7	4				100.0
Sig.		.20	.20	.19	.07

(b) eyebrows

		Subset			
Pos.	N	1	2	3	4
0	4	0.0			
1	4	3.5			
2	4	11.8	11.8		
3	4	44.5	44.5	44.5	
5	4		62.3	62.3	62.3
4	4			68.6	68.6
6	4			87.5	87.5
7	4				100.0
Sig.		.15	.07	.18	.31

(c) nose

		Subset	
Pos.	N	1	2
0	4	0.0	
1	4	3.3	
2	4	7.6	
3	4	14.6	
6	4		69.5
4	4		73.8
5	4		77.7
7	4		100.0
Sig.		.98	.55

(d) mouth

		Subset		
Pos.	N	1	2	3
0	4	0.0		
1	4	11.9	11.9	
2	4	17.7	17.7	
4	4	24.5	24.5	
3	4	26.6	26.6	
5	4		48.7	
6	4			94.9
7	4			100.0
Sig.		.43	.11	1.00

(e) hair

		Subset		
Pos.	N	1	2	3
0	4	0.0		
2	4	16.9	16.9	
1	4	18.8	18.8	
3	4	34.9	34.9	
5	4		51.7	51.7
4	4		62.0	62.0
6	4			90.9
7	4			100.0
Sig.		.29	.08	.05

(f) outline

		Subset	
Pos.	N	1	2
0	4	0.0	
1	4	6.6	
2	4	21.7	
3	4	22.5	
4	4	28.0	
5	4		82.7
6	4		92.1
7	4		100.0
Sig.		.26	.79

Tukey HSD, means for groups in homogenous subsets are displayed for each of the six features (a) eyes, (b) eyebrows, (c) nose, (d) mouth, (e) hair, (f) outline. Uses Harmonic mean Sample Size = 4, α=.05.

It can be seen that no features show the exact same pattern of effects as that seen in the overall homogenous subsets for the main effect of replacement (Table 5.3-3). Not all patterns of means are completely linear from replacement 0 to 7, such as for the nose and is discussed later in section 5.4 Phase 2 Discussion. However, a commonality between the individual feature subsets shows that no significant difference is observed between replacing features at replacement 5 and 6 (except mouth) and 6 and 7.

Familiarity of targets: To assess if a target's familiarity to participants affected miss rates, the overall mean miss rate (collapsed across replacement and feature) was compared to its familiarity to participants as indicated in the Spontaneous naming task in the Control section (familiar valid trials as a percentage of total trials shown). This would indicate how 'popular' the target face was to the participant pool. Results showed no reliable correlation (*r*(22) = .17, *p* = .43). To assess if this familiarity score impacted on miss rates without collapsing feature and replacement, the familiarity to participants score (%) was used as a covariate in a Univariate ANOVA with the factors of replacement and feature. Results showed that the covariate was found to be significant (*F*(1,107) = 93.91, *p* < .001, η_p^2 = .47). A scatterplot indicated a positive relationship; the higher the familiarity to participant score, the higher the miss rates. This will be interpreted in the Phase 2 Discussion (see 5.4 Phase 2 Discussion).

The familiarity score was also tested against those stimuli that received correct recognition naming for when all features had been replaced (replacement 6) to see establish whether familiarity with the target facilitated recognition when all features had been replaced, which could be considered a difficult task. Results showed no reliable correlation (*r*(22) = .34, *p* = .11). The other covariates of attractiveness, memorability and trustworthiness were also found to not be reliably correlated (*p* > .78).

Explicit Choice: Overt recognition responses to the stimuli were collected in the form of an Explicit choice task (multiple choice, n=3) and compared to the Spontaneous naming task responses, for each item collapsed across feature and replacement conditions. This was to establish if participants who were able to spontaneously name a target face were also likely to be able to overtly recognise the face. Explicit choice miss rates (%), for valid trials only, were compared to miss rates (%) for the Spontaneous naming task. As expected, there was a significant positive correlation (*r*(22) = .68, *p* < .001): the higher the spontaneous miss rate, the higher the explicit choice miss rate. This correlation is expected as it is likely that participants who could correctly recall the name of the target would subsequently choose the correct name in the Explicit choice task.

Lecturer version

Participants: Seventy-one participants (UCLan = 40, LJMU = 31) aged 18 or over (age: *M* = 21.3 [age not declared = 5] range = 18-47 yrs.) completed the online study (males n = 11, females n = 60, other n = 0, not declared n = 3).

Data: Data screening was the same as for the previous phases (see section 4.3 Results) and collapsed across the two institutions. Checks were made for missing data and found twenty stimulus images received no valid trials. This resulted in mean familiar (valid) trials of 2.2 per stimulus (Total trials seen: *M* = 5.9). This meant that 36.9% of all trials were familiar and, therefore, valid for analysis (n = 195 trials (*M* = 32 per replacement)). Low numbers of valid trials were expected as not all lecturers will have been familiar to the participants as they were recruited from different courses.

Table 5.3-5: Phase 2 (lecturer) – Mean miss in percentage by feature and replacement

		feature						
		eyes	nose	outline	mouth	hair	eyebrows	Mean
replacement	1	7.1	0.0	5.6	0.0	7.1	0.0	3.5
	2	50.0	-	5.6	11.1	0.0	8.3	13.9
	3	-	7.1_2	0.0	33.3	42.9	58.9	26.9
	4	49.2	66.1	28.6	0.0_1	85.7	0.0_1	46.2
	5	90.5	28.6	46.4	0.0_2	21.4	57.1_2	45.8
	6	14.3	69.0	64.3	66.7	46.4	50.0	53.1
	Mean	47.6	41.6	28.4	27.1	37.6	36.6	36.2

Mean miss rates (%) of each feature at their replacement replacements, collapsed across items. All mean values contained a minimum of 3 observations, unless otherwise stated in subscript. "-" denotes missing data (2 cells).

Descriptive analysis: Table 5.3-5 shows a mostly ordinal increase for the overall mean of replacement (see last column), similar to the pattern of effects for the celebrity data, except that replacement 5 yields a slightly lower mean than replacement 4. Outline follows a mostly ordinal pattern (which is difficult to assess anyway due to missing data); the remaining features do not, as can be seen at replacement 6 for eyes and replacement 5 for hair. Miss rates were generally very low at replacement 1 and quite high at replacement 6. Overall mean miss rates (last row) were fairly variable across features (range = 27.1-47.6%). The range is larger here (*MD* = 20.5%), compared to the celebrity data (*MD* = 11.5%). This would be expected since there are fewer data points here and so estimates are likely to be less accurate. The highest miss rate across replacements was for eyes, followed by nose, hair, eyebrows, outline and mouth. For replacement 6, it was hypothesised that replacing all six features would result in 100% miss rates. However, the total miss rate (not overall Mean) for replacement 6 was 80.7%. Across all valid trials, 12.9% of responses were recorded as 'familiar'; 6.5% were correct recognitions (n = 2 responses across two targets).

Analysis of Variance: Repeating the celebrity analysis methodology, a baseline score of zero and ceiling score of 100 was used for items (8 levels) (see section 5 Celebrity version – analysis of variance, for more detail). For the full model, Univariate ANOVA for the factors of feature (EB,N,O,M,H,E) and replacement (0-7). There were too few cases for males (total = 50, *M* = 1.4) and so the factor of target sex was not analysed.

Univariate ANOVA for the factors of feature and replacement revealed a significant main effect of replacement (*F(*7,48) = 27.97, *p* < .001, η_p^2 = .80) but no significant main effect of feature (*F(*5,48) = 1.17, *p* = .34, η_p^2 = .11). The homogenous subsets of replacement (Tukey HSD post-hoc tests groups levels together that are not significantly different from one another) shown in Table 5.3-6, show replacements split into three groups.

Table 5.3-6: Phase 2 (lecturer) - Homogenous subsets of replacement

replacement	N	Subset		
		1	2	3
0	12	0.0		
1	13	3.5		
2	12	13.9		
3	9	26.9	26.9	
5	12		45.8	
4	11		46.2	
6	13		53.1	
7	12			100.0
Sig.		.06	.07	1.00

Tukey HSD, means for groups in homogeneous subsets are displayed (Between-subjects factor of feature and replacement). Based on observed means and Alpha = .05. Results show differences between feature replacement in three groups. Note that replacements 4 and 5 appear in a non-ordinal order.

This time, at least four features need to be replaced before a significant increase in miss rates between the next feature replacement. However, it is the 5th feature being replaced that is driving this (replacements 4 and 5 are not in numeric order): the means of replacements 4 and 5 are very similar. Further feature replacement is grouped together, except for between replacement 6 and 7. Cohen's d was calculated from one subset to the next and found both differences (Subset 1-2, 2-3) showed a large effect size (d > 1.21). Cohen's d for replacements showed medium effect sizes between the first four feature replacements (d > 0.49), small effect sizes for two further replacements (d < 0.22) and a large effect size from replacement 6 to 7 (d = 1.67) (overall mean, d = 0.59).

Simple main effects: A significant interaction between feature and replacement was found ($F(33,48) = 2.10$, $p = .009$, $\eta_p^2 = .59$). Simple main effects were conducted for feature and replacement. For the effect of feature replaced on replacement, an approaching effect was found for replacement 2 (p = .069). No other replacements revealed any significant effects ($p > .12$). However, all features (except eyebrows) showed a significant effect of replacement ($p < .012$). Eyebrows showed an approaching significant effect of replacement ($p = .05$). Tukey HSD post-hoc homogenous subsets could not be created for each feature due to one or more levels receiving less than two cases.

Item familiarity: To assess a target's familiarity, the overall mean miss rate (collapsed across replacement/feature) was compared to its familiarity to participants as indicated in the Spontaneous naming task in the Control section (familiar valid trials as a percentage of total trials shown). Results showed no reliable correlation (Pearson's *r*(15) = 0.17, *p* = .55).

The item familiarity score was also tested against correct recognition naming scores for when all features had been replaced (replacement 6). Results showed no reliable correlation (*r*(13) = .30, *p* = .28).

Explicit Choice: Explicit choice miss rates (%) for valid trials only were compared to miss rates for the Spontaneous naming task and as expected there was a significant positive correlation (*r*(12) = .61, *p* = .021) (one missing data point with no valid trials). the higher the spontaneous miss rate, the higher the explicit choice miss rate.

Combined

For Phase 2, celebrity (24 cases) and lecturer (15 cases) data were able to be combined to give a broader and more general understanding of the effect of replacing target features.

Table 5.3-7: Phase 2 (celebrity and lecturer) Mean miss in percentage by replacement by individual feature

		feature						
		eyebrows	nose	outline	mouth	hair	eyes	Mean
replacement	1	2.3	2.2	6.1	7.9	14.9	6.3	6.6
	2	10.6	7.6_4	14.8	14.9	11.3	34.8	16.0
	3	49.3	12.1	15.0	28.9	36.5	54.2_4	31.2
	4	54.9	71.2	28.2	19.6	69.9	59.5	51.6
	5	60.6	67.9	70.6	32.5	41.6	84.4	60.1
	6	75.0	69.3	82.8	85.5	76.1	59.1	74.5
	Mean	43.9	42.8	38.5	36.1	43.9	50.5	42.6

Means calculated across both versions. Overall mean miss rates for feature (bottom row) and replacement (last column). All mean values contained at least 5 observations except two points that contained 4, in subscript.

Descriptive analysis: Table 5.3-7 shows an ordinal increase from replacement 1 to replacement 6 (last column). An ordinal increase can also be observed for eyebrows and outline; the other features do not show this same pattern. However, most show a low miss rate for replacement 1 and a high miss rate for replacement 6.

Analysis of Variance: Univariate ANOVA analysis of variance for the factors of feature (EB,N,O,M,H,E) and replacement (0-7) was carried out. As expected significant main effects were found for feature ($F(5,238) = 2.30$, $p = .046$, $\eta_p^2 = .046$) and replacement ($F(7,238) = 100.89$, $p < .001$, $\eta_p^2 = .75$). Tukey HSD post-hoc comparisons found the main effect of feature, could in part, be due to a significant difference between eyes (highest miss rate) and mouth (lowest) ($p = .011$) and an approaching significant difference between eyes and outline (fifth) ($p = .056$). To assess the pattern of miss rates across feature replacement order (0-7), polynomial contrasts for the factor of replacement yielded a linear and quadratic pattern of results ($p < .004$). This pattern aligns with the hypothesis that there would be an ordinal relationship between each replacement position as well as a potential exponential increase in miss rates once 'n' number of features had been replaced, when collapsed across feature. The homogenous subsets of replacement shown in Table 5.3-8, show replacements split into six groups.

Table 5.3-8: Phase 2 (Combined) - Homogenous subsets of replacement

replacement	N	Subset 1	2	3	4	5	6
0	36	0.0					
1	37	6.6	6.6				
2	36		16.0				
3	33			31.2			
4	35				51.6		
5	36				60.1	60.1	
6	37					74.5	
7	36						100.0
Sig.		.89	.55	1.00	.68	.08	1.00

Tukey HSD, means for groups in homogeneous subsets are displayed (Between-subjects factors of feature and replacement). Based on observed means and Alpha = .05. Results show differences between feature replacement in six groups.

At the start, at least two features need to be replaced before a significant difference occurs in miss rates between the next feature replacement. Further feature replacement is paired. Cohen's d was calculated from one subset to the next (Subset 1-2, 2-3, 3-4, 4-5, 5-6) and found a large effect size between subsets 2 to 3, 3 to 4, and 5 to 6 ($d > 0.80$). A medium effect size was found between subsets 1 and 2 ($d = 0.59$) and a small to medium effect size between 4 and 5 ($d = 0.36$). Cohen's d for replacements showed medium to large effect sizes between the first four feature replacements ($d > 0.63$), a small effect size between replacements 4 and 5 ($d = 0.25$), a medium effect size between replacements 5 and 6 ($d = 0.47$) and a large effect size between 6 and 7 ($d = 0.89$) (overall mean, $d = 0.61$).

Simple main effects: An interaction between replacement and feature was found ($F(35,238) = 2.25$, $p < .001$, $\eta_p^2 = .25$). The effect of feature was found to be significant at replacements 2-5 ($p < .049$) but not for replacement 1 ($p = .314$) and 6 ($p = .672$). Tukey HSD pairwise comparisons only yielded a significant difference between mouth (lowest) and eyes (highest) for replacement 5 ($p = .032$); an approaching significant difference between mouth (lowest) and nose (highest) and hair (2nd) for replacement 4 (p < .10); and between eyes (highest) and nose (lowest), and between eyebrows (5th) and hair (4th) for replacement 2 (p < .08). For replacing eyes at replacement 5, the least salient eyebrows are yet to be replaced. In contrast, for replacing mouth at replacement 5 (lowest) the more salient hair remains to be replaced.

Simple main effects revealed that replacement was significant for all features (p < .001). Post-hoc Tukey HSD Homogenous subsets were calculated for each feature. Given the value of these subsets for interpretation of the results, all six subsets are presented in Table 5.3-9.

Table 5.3-9: Phase 2 (Combined) – Homogenous subsets of replacement by individual feature

(a) eyes

		Subset			
Pos.	N	1	2	3	4
0	6	0.0			
1	6	6.3			
2	6	34.8	34.8		
3	4		54.2	54.2	
6	6		59.1	59.1	
4	7		59.5	59.5	
5	7			84.4	84.4
7	6				100.0
Sig.		.13	.51	.26	.91

(b) eyebrows

		Subset			
Pos.	N	1	2	3	4
0	6	0.0			
1	6	2.3			
2	6	10.6	10.6		
3	6		49.3	49.3	
4	5		54.9	54.9	54.9
5	6			60.6	60.6
6	6			75.0	75.0
7	6				100.0
Sig.		.99	.06	.62	.06

(c) nose

		Subset	
Pos.	N	1	2
0	6	0.0	
1	6	2.2	
2	4	7.6	
3	6	12.1	
5	5		67.9
6	7		69.3
4	6		71.2
7	6		100.0
Sig.		.97	.17

(d) hair

		Subset				
Pos.	N	1	2	3	4	5
0	6	0.0				
2	6	11.3	11.3			
1	6	14.9	14.9			
3	5	36.5	36.5	36.5		
5	6		41.6	41.6	41.6	
4	6			69.9	69.9	69.9
6	6				76.1	76.1
7	6					100.0
Sig.		.08	.23	.14	.12	.24

(e) mouth

		Subset	
Pos.	N	1	2
0	6	0.0	
1	6	7.9	
2	7	14.9	
4	5	19.6	
3	6	28.9	
5	6	32.5	
6	6		85.5
7	6		100.0
Sig.		.11	.91

(f) outline

		Subset	
Pos.	N	1	2
0	6	0.0	
1	7	6.1	
2	7	14.8	
3	6	15.0	
4	6	28.2	
5	6		70.6
6	6		82.8
7	6		100.0
Sig.		.16	.13

Tukey HSD, means for groups in homogenous subsets are displayed for each of the six features (a) eyes, (b) eyebrows, (c) nose, (d) mouth, (e) hair, (f) outline. Uses Harmonic mean Sample Size = 4, α=.05. (Between-subjects factor of replacement).

It can be seen that no features show the same pattern of effects as that seen in the overall homogenous subsets for the main effect of replacement (Table 5.3-8) and only eyebrows and outline show an ordinal increase in miss rates with increasing replacement. Not all patterns of means are entirely ordinal from replacement 0 to 7 (for several reasons, including experimental noise and power, as discussed later). This more complex outcome by feature is illustrated by the polynomial contrasts: contrasts yielded significant linear patterns for all features (*p* < .001), but with the additions of a quadratic pattern for mouth (*p* < .001) and outline (*p* = .003), an Order 4 pattern for eyes (*p* = .025), an Order 5 pattern for nose (*p* = .011) and outline (*p* = .019), an Order 6 pattern for hair (*p* = .005) and an order 7 pattern for nose (*p* = .027).

5.4 Phase 2 Discussion

As shown in Phase 1, replacing a single feature yielded low miss rates. Phase 2 was setup to test how much of the target face needed to be replaced in order to conceal identity as well as which features would be best for replacement. A hypothesis was made that a criterion point would be found somewhere around the middle of feature replacement (3rd or 4th feature) because it is at this point that half the face has now been replaced and half is remaining. The composite face effect (Young et al., 1987) found that identification of its respective component halves was low so it seemed sensible that this would also be the case for the current study, except that it was not known if replacing several different parts across the face would alter these results compared to replacing a whole upper or lower half of the face.

Additionally, the design enabled assessment of which features had been replaced and if any were more important than others. Based on the feature saliency hierarchy results from phase 1, it was hypothesized that the most salient features, eyes and hair, would be most impactful to recognition when replaced than lower saliency features, such as nose and eyebrows.

The following themes emerged from the results:

1. Criterion points – Miss rates increased significantly at certain points during feature replacement, suggesting a shift from the original target identity, towards a new one.
2. Familiarity – There were some differences in miss rates and criterion points between the different versions of the experiment (celebrity and lecturer).
3. Feature saliency – Some replaced features were more important to recognition than others.
4. Residual identifiable information – At replacement 6, all target features had been replaced, however, this did not result in 100% miss rates overall.

This chapter will discuss the impact of replacing features incrementally and the potential criterion point found for all data (overall). Following sections will discuss the differences found between celebrity and lecturer data as well as levels of familiarity, feature saliency and possible explanations for residual recognition of stimuli where all six features had been replaced. As before, overarching themes, such as stimulus specific results, potential limitations with the study and contributions to practice and theory, will be discussed in the General Discussion.

Criterions

All data: As expected, an ordinal increase in miss rates, for all data, followed the incremental replacement of facial features, supported by a linear and (sometimes also) quadratic pattern and overall effect of replacement. Miss rates overall were lowest at replacement 1 (*M*=6.6%) and by far the highest at replacement 6 (*M*=74.5%) (see Table 5.3-7). This consistent ordinal increase was also observed when the eyebrows and outline were replaced. However, other features did not show this entirely consistent ordinal pattern, although replacement 1 generally yielded the lowest miss rates (except hair) and replacement 6 the highest (except eyes and nose). It is possible that this lack of monotonic increase in miss rates could be attributed to which features had been replaced beforehand with some having more of an impact depending across different items (see Table 5.2-1). This is supported by a significant effect of feature as well as interaction with replacement. As expected, all features showed an effect of replacement, but only replacements 2-5 showed an effect of which feature was being replaced. Only replacement 5 yielded a significant difference, between the mouth (lowest miss rate) and eyes (highest miss rate): replacing mouth at replacement 5 meant that the more salient hair was yet to be replaced. In contrast, replacing eyes at replacement 5 meant that the low salience eyebrows remained to be replaced. Therefore, it is possible that this result is driven by the feature that is remaining to be replaced and whether that is more salient or not.

In contrast, replacements 1 and 6 were not affected by which feature was being replaced at that point. This could be because one feature replacement has little effect on recognition, as demonstrated in phase 1 where the overall mean miss rates were 9.1%. Additionally, Tukey's HSD homogenous subsets grouped replacement 0 and 1 together, suggesting no significant increase in miss rates when replacing one feature, compared to its veridical self (see Table 5.3-8). For the lack of effect of feature on replacement 6, it is possible that due to a high level of miss rates at replacement 5, any subsequent feature replacement at replacement 6 may not have much of an effect and because of this, feature saliency does not make much difference at this point. This is supported by the homogenous subsets, that groups replacement 5 and 6 together. This analysis also separated the other features into mostly pairs, except replacement 3, giving a total of six subsets and suggests that at least 2 features need to be replaced before the first significant increase in miss rates. The combined data gave the most powerful statistical analysis, compared to the individual versions (celebrity and lecturer), that has resulted in a significant increase in miss rates quite early in feature replacement (replacement 2). This is supported by large effect sizes from positions 0-2. It is also evident that most power is likely to have been around the mid-point of the scale/range, as demonstrated by replacement 3 being significantly different from replacements 2 and 4 (only one change needed in the middle of feature replacement).

None of the homogenous subsets for the individual features showed the exact same pattern as the overall results, in terms of their groupings (see Table 5.3-9). One possible explanation for this is that feature replacement may be affected by which features have been replaced before. For example, replacing the eyebrows at replacement 2 meant that the more salient eyes had already been replaced at replacement 1, therefore the miss rates may already be high and the subsequent lower saliency eyebrow replacement did not affect miss rates as much, grouping those two replacements together. Additionally, only eyebrows and outline showed ordinal increases in miss rates, the other features did not. It would be expected that a further feature replacement would always result in either equal or higher miss rates, not less. One possible explanation for this is the difference in the effect of the feature rotation order on different stimuli. Six rotation orders were used, with items split between them, and therefore within each feature subset table, each replacement result was yielded by a different set of items. For example, replacement 4 yielded higher miss rates than replacement 6, for the eyes. The results for eyes at replacement 4, came from rotation B (7 items), and replacement 6 came from rotation C (7 different items). It is possible that the rotation B items were more affected by the four feature replacements, than the six feature replacements for the rotation C group, resulting in non-ordinal increases in miss rates. Therefore, it seems that the effect of the quantity of feature replacement is not always even across different stimuli, with some being more affected than others. Additionally, for the nose, replacements 4-6 are very similar for miss rates and so changes in the ordinal order would be understandable. The overall homogenous subset (see Table 5.3-8) averages across these inconsistencies to give a consistently ordinal result. Even so, all features showed a linear contrast pattern of distribution with additional quadratic patterns for the mouth and outline, as well as further patterns for some other features, suggesting that although distributions were broadly linear, there was some understandable deviation from this pattern.

Familiarity

Differences were found in both the amplitude and homogenous subset results between the celebrity and lecturer versions suggesting that the type of familiarity has an effect on the results, both quantitatively and qualitatively.

Performance: Overall mean miss rates were moderate for both versions with the celebrity (45.7%) miss rates somewhat higher than for lecturers (36.2%). One possible explanation for this is a difference in level of familiarity, as discussed in phase 1 (see section 2.2 Familiar and unfamiliar face recognition for a literature review). However, both versions, like the combined data, showed an overall ordinal increase in miss rates by replacement (exception replacements 4 and 5 for lecturers) that is supported by linear patterns of distributions and a significant effect of replacement. This supports the prediction that the less information remaining in the face, the less chance there is of the face being recognised. Homogenous subsets suggest that at least four features need to be replaced for a significant increase in miss rates, but this result is skewed by the non-ordinal results of replacements 4 and 5. The slightly later criterion point for the lecturer data (replacement 4) compared to the celebrity (replacement 3) could be attributed to the generally lower miss rates for lecturer data that may not have allowed for a significant change until during later stages of feature replacement. Effect sizes (Cohen's D) for position also reflected a difference in the pattern of results with the mean effect size by position higher for celebrities (*d* = 0.68) than lecturers (*d* = 0.59). This supports the finding that replacing features has more of an effect on recognition of celebrity faces, compared to lecturers.

In comparison, the criterion points found differed between target type and for combined data, where the first significant increase in miss rates occurred upon replacing at least two features, three features for celebrity data and four for lecturers. However, this theoretical criterion point only marks a point at which miss rates increase significantly: it does not mark a point at which identity is concealed for all participants, as this point was never found. It should also be noted that the celebrity data feature hierarchy in phase 1i suggested a preference towards the internal features as an indication of higher familiarity, compared to the lecturer version that indicated an external feature preference. One would also expect that if a face is less familiar that the miss rates will be higher because of featural manipulations. However, the miss rates were higher for celebrities compared to lecturers (*MD* = 9.5%). This was in contrast to phase 1i where celebrities yielded lower miss rates than lecturers (*MD* = 12.0%). This mirror effect could be explained by the difference in the impact of featural disruptions to the holistic whole-face context effect (Tanaka and Farah, 1993), which is higher for more familiar faces (celebrities): phase 1i only replaced one feature and was less disruptive to the whole-face context (lower miss rates for celebrities than lecturers) compared to the compound feature replacement of phase 2 that was more disruptive and, therefore, had a greater impact on recognition for more familiar faces (celebrities higher miss rates than lecturers) (see section 2.4 Holistic processing, for a literature review on holistic processing). The homogenous subsets for replacement found that a significant increase in miss rates (criterion point) occurred later for the lecturer data (replacement 4) compared to celebrities (replacement 3). This is surprising given that one would expect a less familiar face to yield significant increases in miss rates earlier on in feature replacement. However, the lecturer data yielded fewer responses per item, which have likely increased noise and made the power of the test lower for lecturers (this is evidence by much higher standard deviation scores for the lecturer data, compared to celebrities). Given this point, it is perhaps more ecologically valid to use the criterion points for the celebrity data that have higher numbers of responses per item and represent a higher level of familiarity. The celebrity data suggests that at least three features (half the face) need to be replaced before a significant increase in miss rates.

It should be noted that celebrity data was initially analysed using an independent subjects design for the between subjects factors of feature and replacement, followed by a subsequent more powerful analysis using a mixed-factorial design where feature was able to be analysed as a within subjects factor (this could not be carried out for the lecturer data due to missing data points, so only independent subjects analysis was carried out). Homogenous subsets for the celebrity data were generated for both analyses with slight differences between the two: replacement 3 shifts from group two to group one. This difference could be attributed to a difference in how the error term was calculated (it is lower for feature in the mixed factorial design, giving more power). Reported above, are the subsets for the mixed-factorial analysis.

Trials: The number of valid trials for the celebrity version was quite high at 85.1% of the total trials and in contrast, the lecturer version of the experiment yielded valid trials of only 36.9% of the total trials shown. The reasons for this are the same as for phase 1 and caused by differences in recruitment and participant pools for the lecturer version (see Phase 1 section 4.6 for a discussion on Familiarity). Recruitment for the lecturer version was low at seventy-one participants compared to ninety-one for the celebrity version.

The lower lecturer numbers were as a result of a logistical timing issue, with data collection only being able to be completed mainly over the second academic semester.

Similar to phase 1, further analysis was carried out to assess whether the likelihood of familiarity for each target had an effect on miss rates. Both phase 2 experiments showed no significant correlation between miss rates and the level of familiarity per target. This may suggest that a target's popularity (or level of familiarity) does not affect the recognition rates under the stimulus manipulations. Unlike phase 1, which did find a correlation, and even though there were clear quantitative and some qualitative differences in miss rates between the different versions of the experiment, this does not appear to be caused by the level of familiarity, or popularity, indicated through the number of valid trials. However, further analysis showed that when the familiarity score was used as a covariate in an ANOVA for replacement and feature for the celebrity data, it had a significant effect: the higher the familiarity to participant score, the higher the miss rates. This is in contrast to the expectation the familiarity would render a target more invariant to featural manipulations, but it seems that these manipulations are more disruptive for more popular compared to less popular faces, via holistic processing effects. This popularity score seems to be providing some stimulus specific effect under the various featural manipulations, suggesting that in this context, the familiarity of a target face may change how it is affected by feature replacement and how many features have been replaced, but not for the overall miss rate.

Feature saliency

Both celebrity and lecturer data yielded no significant main effect of feature, even though when combined, a main effect of feature was found. However, this was only due to a significant difference between the eyes and mouth and an approaching significant difference with the outline. This suggests a slightly higher saliency for when the eyes are replaced across all replacements, which aligns with previous literature (see section 2.3 Feature saliency hierarchy). As indicated by the individual homogenous subsets, the only features to also show an ordinal increase in miss rates for replacement were outline and eyebrows (for combined data) and eyes and outline for the celebrity data with none for the lecturers, which was not able to be assessed in this way due to missing data. Outline showed a consistent result between the celebrity and combined data. However, miss rates were generally low at replacement 1 and high at replacement 6 for all features for celebrities and lecturers, with similar overall miss rates across features, supporting the lack of overall effect of feature in these individual versions. Both versions did find an interaction between replacement and feature, even when no main effect of feature was found. For the celebrity data, this interaction was attributed to an effect of feature at replacement 4 only. Further analysis using contrasts found that there was a significant difference between the mouth (lowest miss) and nose (highest miss) at this replacement. Upon further inspection of the data, the mouth followed the replacement of less salient features (eyebrows, nose and outline) compared to the nose which followed the replacement of high saliency features (eyes, hair and less salient eyebrows). Therefore, it seems that, in general, which features have already been replaced does not make a difference to feature replacement, except at replacement 4, which coincidentally is also just after the point at which a significant increase in miss rates occurred (replacement 3). From this, it seems that the feature being replaced, as well as which features have been replaced prior, are not important, until the critical point at which the stimulus switched from old to new identity. In this case, the features replaced before the nose could be considered to make up the upper half of the face (nose, eyes, hair and eyebrows). The features replaced after the mouth could be considered to make up mostly the lower half of the face (mouth, nose, outline and eyebrows (not lower face, but least salient)). This result could be likened to the composite face effect (Young et al., 1987). The lecturer data also showed a significant interaction between feature and replacement that showed no effect of feature on all replacements, but a significant effect of replacement for all features, except the eyebrows (although this was approaching).

Homogenous subsets for individual features for the lecturer data could not be produced as some items had less than two responses. However, they were for the celebrity data, which showed that not all features yielded an ordinal increase in miss rates. However, even though there were differences between subsets for the features, there was one commonality: there were no significant differences between replacements 5 and 6 for all features (except the mouth). This suggests that after replacement 5, further feature replacement of the sixth feature does not make much difference to miss rates and that even further replacement to replacement 7 (100% miss rate, or different identity) also does not yield a significant difference.

Residual information

One key finding from the results of phase 2 demonstrated that recognition was still possible for familiar faces even when all six features had been replaced. Replacement 6 stimuli had all six features replaced so that there were no remaining target features. Figure 5.4-1 shows the celebrity and lecturer targets that received correct recognitions at replacement 6 in phase 2, accounting for thirteen correct recognitions across six targets.

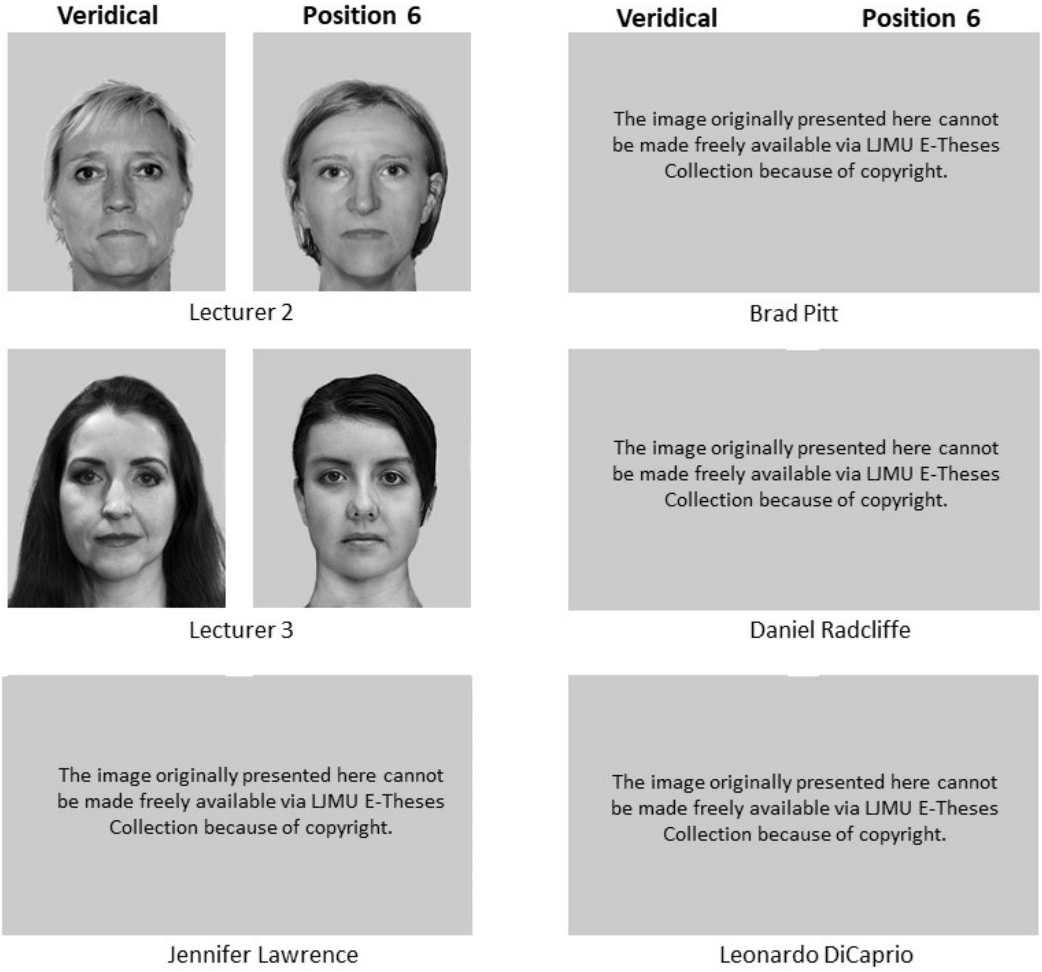

Veridical
Position 6
Veridical
Position 6
The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.
Lecturer 2
Brad Pitt
The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.
Lecturer 3
Daniel Radcliffe
The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.
The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.
Jennifer Lawrence
Leonardo DiCaprio

Figure 5.4-1: Targets receiving correct recognition for replacement 6

Six targets received correct recognition responses for when their respective composite with all six features replaced, was shown. Paired images for targets (Left: Veridical image; right: replacement 6 composite) (Number of Correct recognitions; lecturer 2=1, lecturer 3=1, Jennifer Lawrence=1, Brad Pitt =6, Daniel Radcliffe=3, Leonardo DiCaprio=1).

This result is supported by the differences in miss rates between replacement 6 (74.5%) and replacement 7 (100.0%, everything changed) which was significant for the combined data, suggesting that there was still substantial information available for recognition (*MD*=25.5%) between the miss rate for replacements 6 and 7 (see Table 5.3-8). It should be noted that four of the targets received only one correct recognition. However, Daniel Radcliffe received three correct recognitions (valid trials =12), and Brad Pitt six (valid trials = 14). Both of these figures are fairly high, and therefore unlikely to have occurred by chance, an explanation that might be valid for the other targets who received one correct recognition.

The compositing technique was designed to reduce the manipulations to only featural, by keeping the configuration and contrast of the target constant with the least amount of disruption. Therefore, for replacement 6, the only remaining target information in the image, is the configuration and general contrast of the target. It seems that participants were still able to recognise the targets fairly well with these minimal cues (see section 2.4 Holistic processing for a review on the different streams of information used for recognition). However, it should be noted that with the compositing technique there appears to have been a bias towards the composited features being manipulated in such a way that they would be visually similar to the target in order to generate a plausible face. This can be seen in the composite of Brad Pitt, where the angle of the jawline is the same as the target image and the angle of the eyebrows kept the same to maintain configuration and contrast patterns. This suggests that stimulus specific compositing may account for some of the correct recognitions, where the featural change has been rendered minimal by overly restricting disruption to the configural/contrast streams of information. It is also possible that some target faces were more invariant to feature replacement because of more memorable configurations and/or contrast patterns. Correlational analysis between the three characteristic ratings scores (celebrity data) and the miss rate for replacement 6 revealed no significant correlations, suggesting facial characteristics do not account for this effect. Additionally, further analysis using the likelihood of familiarity score (percentage of total trials that were valid for analysis may indicate some likelihood of familiarity or popularity of the target) also showed no significant correlations when compared to the mean miss rate per target. It is also important to note that this result may be due to some participants having superior face recognition abilities that are still able to recognise faces that have been heavily edited (see Appendix A – Methodological considerations for more detail on face recognition ability). Therefore, it is likely that this result is not due to a stimulus specific effect but rather due to:

a) the compositing technique bias towards matching sampled features to the target in a more congruent way to minimise disruption to the other streams of information.

b) residual configural information and contrast patterns being sufficient for a correct recognition.

c) Some participants having superior face recognition abilities that make them more invariant to featural manipulations.

Summary

When celebrity and lecturer data were combined, the results for phase 2 showed an ordinal increase in miss rates by feature replacement. Replacement also yielded a linear pattern of results as expected, as well as a quadratic pattern of results that seems to demonstrate a criterion point in the replacement order. Homogenous subsets suggest that a significant increase in miss rates, or criterion point, occurred after replacing two features (supported by a large effect size found between position 1 and 2). The celebrity and lecturer versions yielded similar, mostly ordinal, patterns of results for replacement. Feature yielded a main effect for the combined data, where replacing the eyes yielded significantly more miss rates than when replacing the mouth. Feature was also found to have more of an effect in the middle of feature replacement (replacements 2-5) but not for replacements 1 and 6, suggesting feature becomes more impactful around the criterion point where miss rates markedly increase. An effect of feature was not found for the different data types. However, an interaction was found at replacement 4 for celebrities, where prior feature replacement had a significant effect on miss rates. Further analysis showed this to be a difference between the features of the upper and lower halves of the face. When all six features were replaced at replacement 6, recognition was still possible, to some extent, perhaps using residual information or due to participants having superior face recognition abilities.

6 Automated face recognition systems

6.1 Introduction

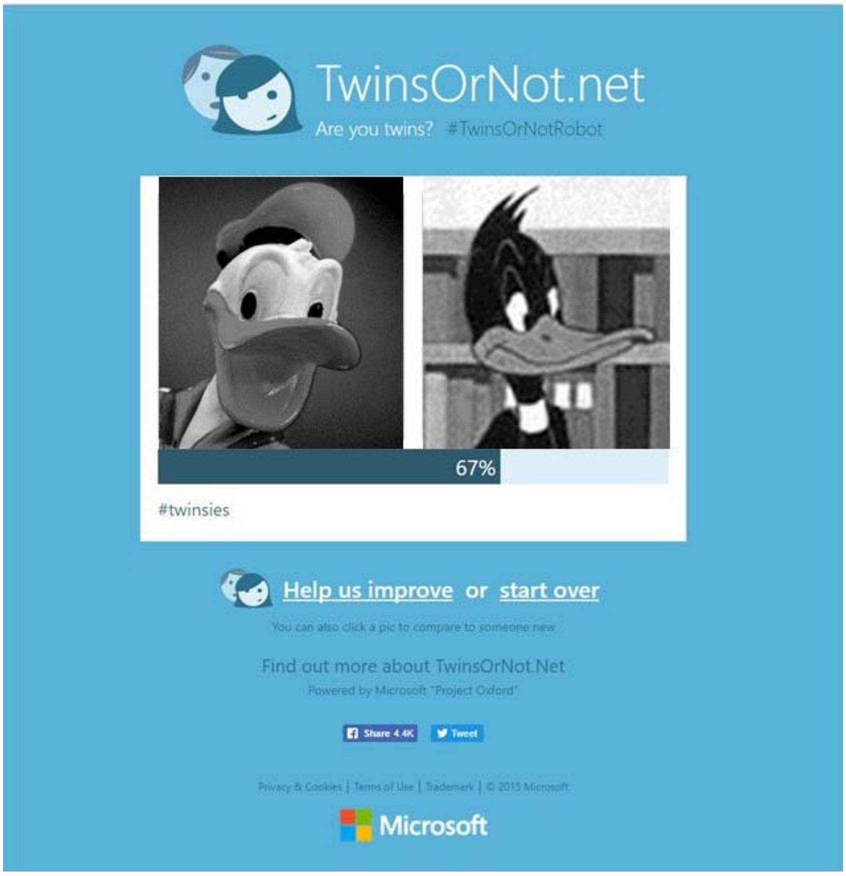
As a comparative way to assess whether the resulting stimuli were convincing to a computer system, the celebrity face stimuli set was run through three automated face recognition (AFR) systems as an ancillary study (see section 2.6 Automated Face recognition systems for a review). It is possible that offenders may check face images for identity using automated systems. The results allowed for the comparison of the three systems tested as well as a comparison with the participant data performance from Phases 1 and 2. Not all automated face recognition systems were available for the public to use, and so a subset of those available and accessible were utilised: Google Picasa (was available during project but is now defunct); the face recognition function in the image editing software, Adobe Lightroom; and the online resource, Twinsornot. As publicly available systems, they will also be available to offenders to check identity of facial images.

6.2 Method

The three automated face recognition systems were adopted and used to test the composite stimulus set from the experimental phases. The methodology for each system is summarised below:

1 Twinsornot: The online automated face comparison system, Twinsornot (www.twinsornot.net) was used to assess how similar the stimuli (under varying feature conditions) were to their respective veridical images. Methodology consisted of pairwise percentage similarity ratings between two images; the veridical target image versus a feature condition stimulus (see Figure 6.2-1).

Images were uploaded to the website temporarily for Twinsornot to make similarity calculations.


TwinsOrNot.net
Are you twins? #TwinsOrNotRobot
67%
#twinsies
Help us improve or start over
Share 4.4K
Tweet
Privacy & Cookies | Terms of Use | Trademark | © 2015 Microsoft
Microsoft

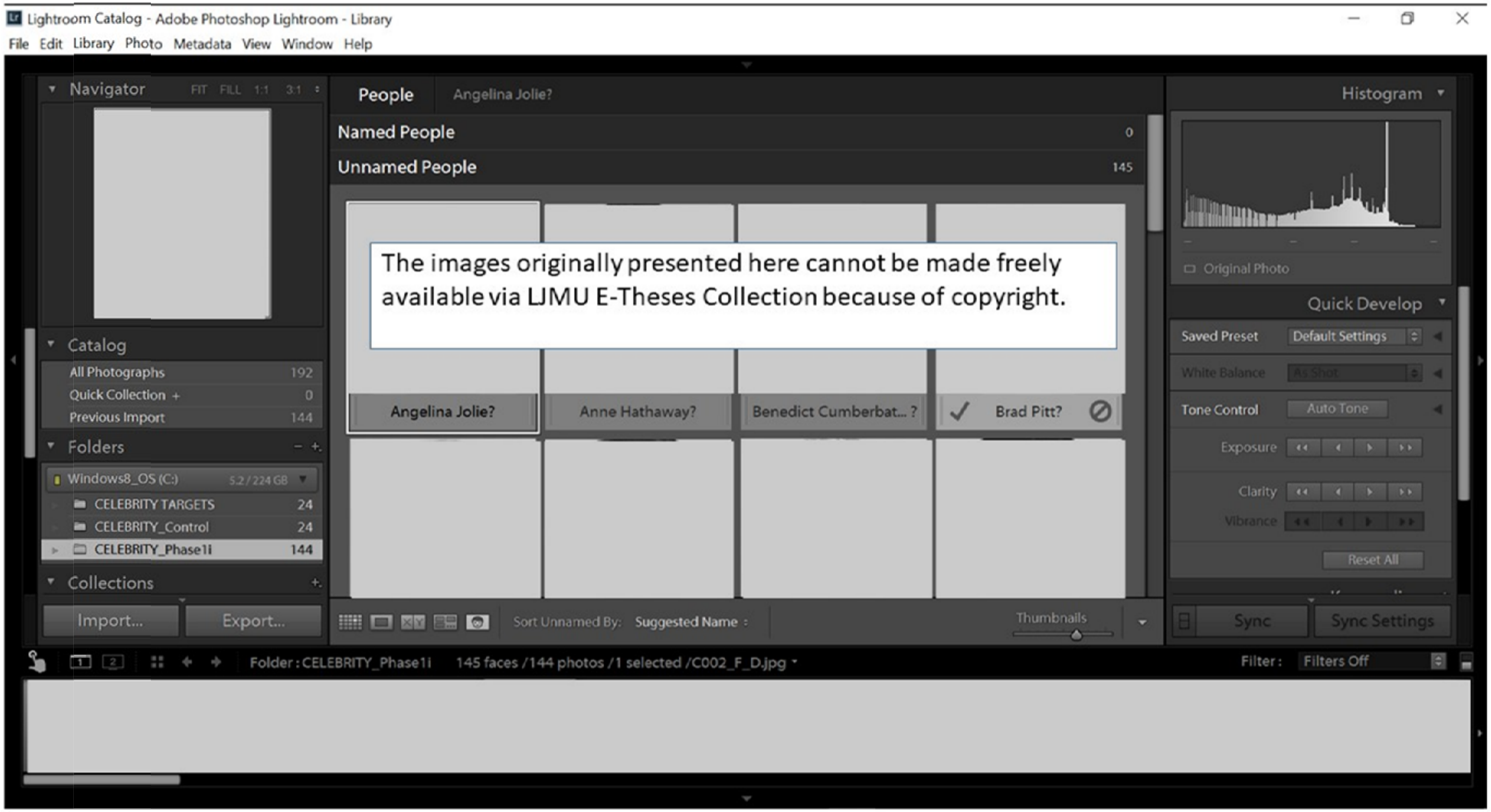
You can also click a pic to compare to someone new

Find out more about TwinsOrNot.Net

Powered by Microsoft "Project Oxford"

Figure 6.2-1: Screenshot of the Twinsornot online GUI

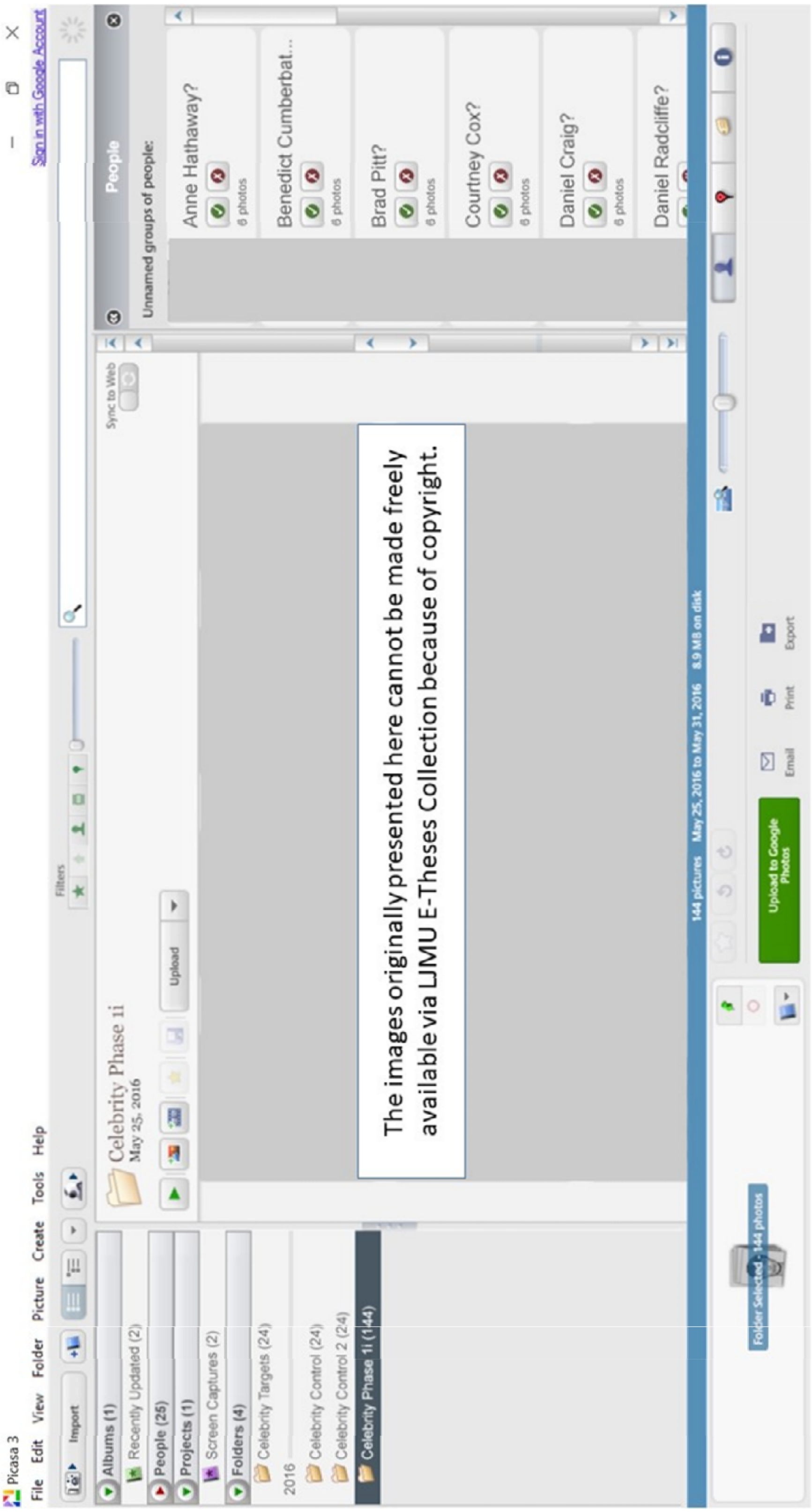
2 Adobe Lightroom: This desktop image editing software was tested to see if the system could recognise the target faces from the experimental stimuli. Veridical images were first uploaded with name tags so that the system learned who the face belonged to. Subsequently the experimental stimuli were uploaded by condition (so that only one image per target was visible at any one time) and where the system recognised the face, suggested a name (see Figure 6.2-2).



The images originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.

Figure 6.2-2: Screenshot of Adobe Lightroom GUI

3 Google Picasa: This desktop photo editing software was utilised in the same way as for Adobe Lightroom (see Figure 6.2-3).



The images originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright.

Figure 6.2-3: Screenshot of the Google Picasa GUI

All systems were tested on a Lenovo ThinkPad (Windows 10) laptop using the installed software (Lightroom and Picasa) and the Internet Explorer browser to access Twinsornot online. Results were recorded in a Microsoft Excel worksheet.

6.3 Results

Phase 1

First analysis: In the first analysis, 144 stimulus images (24 celebrity targets x 6 feature conditions) were run through the three different automated face recognition systems and compared to the veridical target images (n = 24). Results for the three systems for analysis of the Phase 1i stimuli were not vastly different to each other with the results for Picasa and Adobe Lightroom scored as a correct recognition percentage and the scores for Twinsornot as a similarity percentage. For the purposes of analysis, all correct recognition scores were inverted to yield miss rates and the Twinsornot results were also inverted to yield a dissimilarity percentage. Although Twinsornot is a dissimilarity percentage and the other two are miss rates they are being considered comparable for this analysis. Overall mean miss/dissimilarity scores varied between Picasa and the other two systems with Picasa yielding the highest miss rates (38.2%) and both Adobe Lightroom and Twinsornot yielding floor results of 0.0%. Therefore inferential analysis was only carried out on Picasa for the first analysis.

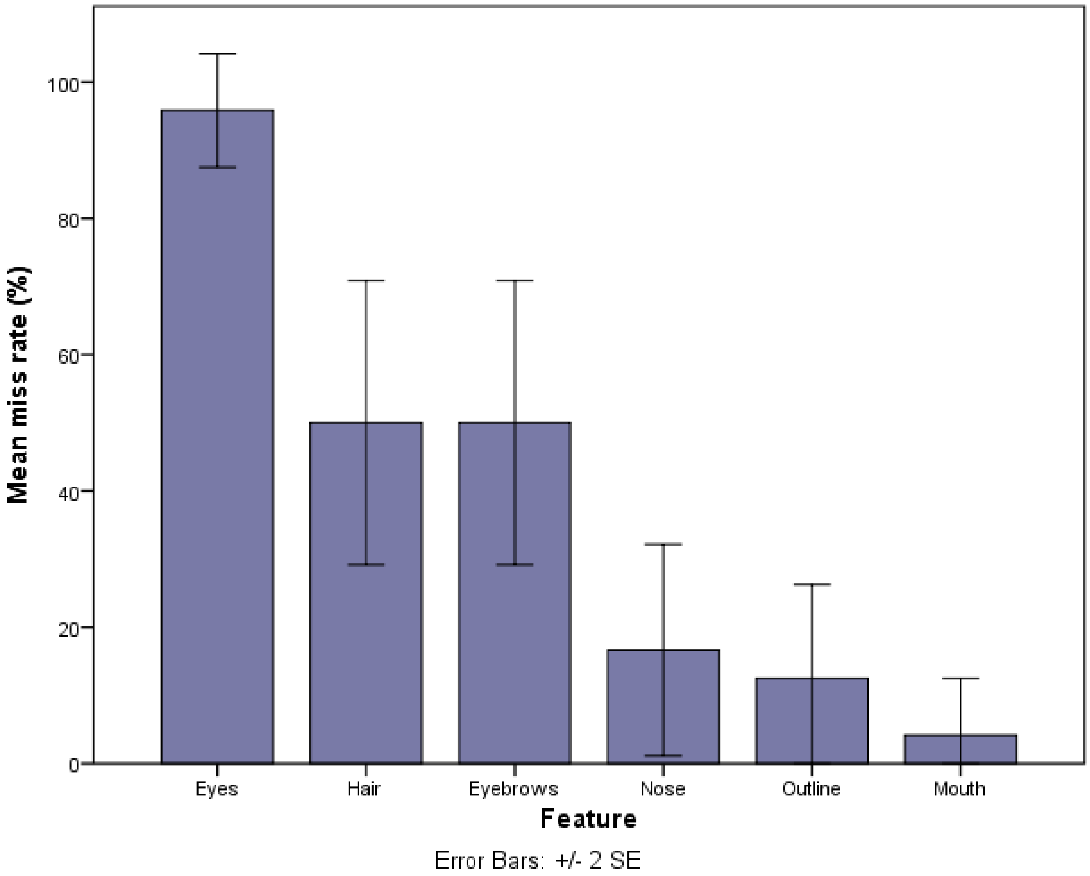


Figure 6.3-1: Picasa – Mean miss in percentage by feature.

Descriptive analysis: Figure 6.3-1 reveals miss rates by feature for Picasa. Results showed the lowest miss rate for mouth (*M* = 4.2%, *SD* = 20.4%) similar to outline (*M* = 12.5%, *SD* = 33.8%) and nose (*M* = 16.7%, *SD* = 38.1%). Eyebrows and hair (both *M* = 50.0%, *SD* = 51.1%) showed much higher and very similar miss rates, with eyes considerably higher, at almost ceiling level (*M* = 95.8%, *SD* = 20.4%).

Analysis of Variance: To test for the overall effect of feature on miss rates, an analysis of variance (ANOVA) was carried out. For Picasa, RM ANOVA for the within-subjects factor of feature (E,EB,N,M,H,O) found a significant main effect of feature (*F(*2.26, 52.04) = 23.68, *p* < .001, η_p^2 = .51). Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated (X_2(14) = 66.36, *p* < .001), and therefore a Greenhouse-Geisser estimate was used (ε = .45). To test if any feature miss rates were significantly different from one another, pairwise comparisons (Bonferroni corrected) were used. Results showed significant differences between eyes (highest miss rate) and all other features (*p* < .014), between mouth (lowest) and eyebrows (3rd) and hair (2nd) (*p* < .014) and approaching significant differences between hair (2nd) and outline (5th) (*p* = .061) and between eyebrows (3rd) and outline (5th) (*p* = .061).

Second analysis: In the first analysis the systems were trained on the original veridical images of the targets from which the composites were generated, therefore it is likely that the composite images contained image cues that the face recognition systems may use for picture matching instead of face recognition (Bruce and Young, 1986). To investigate whether the systems had based their results on pictorial image qualities to help match between veridical target and composite, a second analysis was conducted using three new veridical images of the targets (the original target image was not included). This was not repeated for the Twinsornot system due to a change in the system's online process (updates to the website required anti-bot verification as well as the storage of thumbnails of the images uploaded).

Table 6.3-1: Picasa and Lightroom – Mean miss in percentage by feature (comparison of first and second analysis)

	Picasa		Lightroom	
	1st	2nd	1st	2nd
Eyes	95.8	50.0	0.0	25.0
Eyebrows	50.0	70.8	0.0	20.8
Hair	50.0	79.2	0.0	8.3
Nose	16.7	54.2	0.0	16.7
Outline	12.5	54.2	0.0	16.7
Mouth	4.2	54.2	0.0	12.5
Mean	38.2	60.3	0.0	16.7

Mean miss rates (%) for Picasa and Adobe Lightroom. n=24 per system. Column '1st' lists the means for the first analysis using the original veridical target images. Column '2nd' lists the means for the second analysis using three new veridical target images.

Descriptive analysis: Results, shown in Table 6.3-1, show a very large increase in miss rates for the Lightroom system between floor results for the first analysis (1 original image) to the second analysis (3 new images). For the Picasa system, miss rates from the first analysis also increase considerably in the second analysis overall. The feature saliency hierarchy remains much the same except for a change in position of eyes which yielded the lowest miss rate in the second analysis compared to the highest in the first analysis. This feature hierarchy will be compared with the participant data in a later section.

Analysis of Variance: Each system was analysed using RM ANOVA for the within-subjects factor of feature (E,EB,N,M,H,O) using the second analysis data (3 new images). Picasa showed a significant main effect of feature (*F(*5, 115) = 3.98, *p* = .002, η_p^2 = .15). Post-hoc pairwise comparisons revealed an approaching significant difference between hair (highest miss rate) and eyes (lowest) (*p* = .08). Lightroom showed no significant main effect of feature (*F(*5, 115) = 1.32, *p* = .26, η_p^2 = .05).

Comparison of analyses: For Picasa, RM ANOVA for the within-subjects factor of feature (E,EB,N,M,H,O) and between-subjects factor of method of analysis (1st, 2nd) found a significant main effect of feature (*F(*2.25, 103.62) = 16.80, *p* < .001, η_p^2 = .27, X_2(14) = 112.11, *p* < .001, (ε = .45). A between-subjects effect of method of analysis was found to be significant (*F*(1,46) = 5.56, *p* = .023, η_p^2 = .11): the second analysis generally yielded higher miss rates than the first, except for eyes.

There was also a significant interaction between feature and method of analysis (*F*(2.25,103.62) = 14.41, *p* < .001, η_p^2 = .24). For the interaction of feature and method of analysis, previous analyses showed an effect of feature for both the first and second analyses. The effect of method of analysis was analysed for each feature separately. All features revealed a significant effect (*p* < .036) except for the eyebrows (*p* = .15): for eyes, mouth, hair and outline, the first analysis yielded significantly higher miss rates than for the second (*p* < .036); for nose, the first was higher than the second (*p* = .006).

For the Lightroom system, the ANOVA was repeated and did not find a significant main effect of feature (*F(*5,230) = 1.32, *p* = .26, η_p^2 = .28). There was also no significant interaction of feature and method of analysis (*F(*5,230) = 1.32, *p* =. 26, η_p^2 = .28) but a between-subjects effect of method of analysis (*F(*1.46) = 7.46, *p* = .009, η_p^2 = .14). The second method yielded much higher miss rates than the first.

Comparison to humans: The automated systems data was subsequently compared to the results from the human participant data (celebrity). Data from the second analysis was used as learning from multiple images of the celebrities is more aligned with the way that humans will have encoded a memory from multiple images (i.e. not via picture matching).

Table 6.3-2: Participants and Automated systems - Mean miss in percentage by feature.

	Participants Mean	Picasa Mean	Lightroom Mean
Eyes	18.9	50.0	25.0
Hair	10.0	79.2	8.3
Mouth	9.2	54.2	12.5
Outline	8.3	54.2	16.7
Nose	5.7	54.2	16.7
Eyebrows	2.6	70.8	20.8
Mean	9.1	60.4	16.7

Descriptive analysis: Table 6.3-2 shows that participants performed better at face recognition on every feature change, yielding the lowest mean miss rates overall compared to the two automated systems. The feature saliency hierarchy observed in the participant data is not the same in the automated systems with some main differences: Eyebrows, the lowest saliency feature for participants, yielded the second highest miss rates in the Picasa data and Lightroom; Eyes, the highest saliency feature for participants, is found to be the lowest for Picasa, but congruently the highest for Lightroom; Hair, the second highest saliency feature for participants is comparably high for the Picasa data (first) but the least salient for Lightroom.

Analysis of Variance: RM ANOVA for the within-subjects factor of feature (E,EB,N,M,H,O) and the between-subjects factor of system (participants, Picasa, Lightroom) did not find a significant main effect of feature (*F(*3.47, 239.18) = 1.56, *p* = .19, η_p^2 = .02, X_2(14) = 85.40, *p* < .001, ε = .69). A between-subjects effect of system was found to be significant (*F*(2,69) = 21.15, *p* < .001, η_p^2 = .38). To test if any system performed better or worst than the others, Tukey's HSD post-hoc comparisons were used. Results revealed that Picasa yielded significantly higher miss rates than both other systems (*p* < .001) but not between Lightroom and participants (*p* = .65).

Simple main effects: There was a significant interaction between feature and system (*F*(6.93,239.18) = 3.93, *p* < .001, η_p^2 = .10). Previous analysis showed an effect of feature for participants and Picasa, but not for Lightroom. The effect of system was analysed for each feature separately. All features showed an effect of the system used for analysis (*p* < .023). Tukey's HSD post-hoc tests revealed that the effect of system for each feature was due to Picasa yielding significantly higher miss rates than the other two systems (*p* < .003) (except for eyes, where Picasa was only significantly higher than participants (*p* = .022)).

Phase 1ii

For phase 1ii celebrities (images of single facial features), both Picasa and Lightroom failed to find names that matched to the stimuli. It seems that the systems were unable to process them as they did not appear as full faces. Twinsornot was also unable to detect a face and was therefore unable to output similarity ratings.

Phase 2

First analysis: All of the 144 stimulus images in phase 2 (24 celebrity targets x 6 replacements (feature replaced at that point) presented to the same three automated face recognition systems using the same methodology as for phase 1 (again, inverting the results data to miss rates and dissimilarity scores) (see section 6 Phase 1). As no main effect of feature was found in the celebrity participant data, only replacement was analysed. Overall mean scores for replacement were similar across the systems with Picasa yielding the highest miss rate, moderately more than Adobe Lightroom, which was similar to Twinsornot. See Table 6.3-3 for a summary of the mean miss scores (%) for Picasa and Lightroom and the mean dissimilarity scores (%) for Twinsornot.

Table 6.3-3: Automated Face recognition systems – Mean miss/dissimilarity in percentage by replacement

		Picasa Miss %		Lightroom Miss %		Twinsornot Dissimilarity %	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Replacement	1	4.2	20.4	0.0	0.0	0.0	0.0
	2	12.5	33.8	0.0	0.0	0.0	0.0
	3	8.3	28.2	0.0	0.0	2.3	8.3
	4	8.3	28.2	0.0	0.0	14.3	15.8
	5	25.0	44.2	25.0	44.2	30.4	17.5
	6	70.8	46.4	75.0	44.2	49.0	21.5
	Mean	21.5	41.2	16.7	37.4	16.0	22.6

Miss rates and standard deviation (S.D.) scores for Picasa and Lightroom and dissimilarity scores for Twinsornot (celebrity). N = 24 per system. Bottom row: overall means.

Descriptive analysis: All systems showed an ordinal increase in miss rates/dissimilarity by feature replacement, except for Picasa where replacement 2 yielded higher miss rates than replacements 3 and 4. The change at replacement 6 also yielded particularly marked increases in miss rates for all systems. However, Lightroom and Twinsornot showed floor effects (*M* = 0.0%) for replacements 1-4 and 1-2, respectively. For replacement 6, it was hypothesised that replacing all six features would result in 100% miss rates. However, Picasa yielded correct recognition for 7 targets and Lightroom yielded 6. Twinsornot could not be analysed this way due to the scaled scoring but one target did yield 100% similarity for replacement 6.

Analysis of Variance: In line with the participant analysis (see 5.3 Results), two levels were added to the factor of replacement: “0” and “100”. Univariate ANOVA for the between-subjects factor of replacement was carried out for each system separately. Picasa showed a significant main effect of replacement (*F(*7, 184) = 35.46, *p* < .001, η_p^2 = .57). Tukey HSD post-hoc tests were used to assess the pattern of replacement miss rates in terms of their distribution and as part of that test, Tukey groups levels together that are not significantly different from one another as part of a homogenous subset output table. If replacement levels are significantly different from one another, they are placed in a separate subset. Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into three subsets (see Table 6.3-4).

Table 6.3-4: Phase 2 - Homogenous subsets for replacement (Picasa)

Replaceme		Subset		
nt	N	1	2	3
0	24	0.0		
1	24	4.2		
3	24	8.3		
4	24	8.3		
2	24	12.5		
5	24	25.0		
6	24		70.8	
7	24			100.0
Sig.		.08	1.00	1.00

A significant difference occurs after replacing six features with a further significant increase between replacement 6 and 7.

Lightroom showed a significant main effect of replacement (*F(*7, 184) = 78.86, *p* < .001, η_p^2 = .75). Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into four subsets (see Table 6.3-5).

Table 6.3-5: Phase 2 - Homogenous subsets for replacement (Lightroom)

Replacem		Subset			
ent	N	1	2	3	4
0	24	0.0			
1	24	0.0			
2	24	0.0			
3	24	0.0			
4	24	0.0			
5	24		25.0		
6	24			75.0	
7	24				100.0
Sig.		1.0	1.0	1.0	1.0

A significant difference occurs after replacing five features. Further significant changes occur for replacements 6 and 7.

Twinsornot showed a significant main effect of replacement (*F(*7, 184) = 220.60, *p* < .001, η_p^2 = .89). Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into five subsets (see Table 6.3-6).

Table 6.3-6: Phase 2 - Homogenous subsets for replacement (Twinsornot)

Replacem		Subset				
ent	N	1	2	3	4	5
0	24	0.0				
1	24	0.0				
2	24	0.0				
3	24	2.3				
4	24		14.3			
5	24			30.4		
6	24				49.0	
7	24					100.0
Sig.		1.00	1.00	1.00	1.00	1.00

A significant increase in miss rates occurs after four features are replaced with further increases occurring after each feature replacement.

Comparing the systems: To compare the systems, Univariate ANOVA for the factors of replacement (0-7) and system (Picasa, Lightroom, Twinsornot) found a significant main effect of replacement (*F(7*,552) = 193.20, *p* < .001, η_p^2 = .71), but no significant effect of system overall (*F*(2,552) = 1.92, *p* = .15, η_p^2 = .01). Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into four subsets (see Table 6.3-7).

Table 6.3-7: Phase 2 - Homogenous subsets for replacement across all systems

Replacement	N	Subset			
		1	2	3	4
0	72	0.0			
1	72	1.4			
3	72	3.6			
2	72	4.2			
4	72	7.5			
5	72		26.8		
6	72			65.0	
7	72				100.0
Sig.		.48	1.00	1.00	1.00

Showing replacements 1-4 grouped together

A significant difference occurs after replacing five features. Further significant changes occur after replacing each of the next two features.

Simple main effects: A significant interaction between replacement and system was also found ($F(14,552) = 1.95$, $p = .020$, $\eta_p^2 = .05$). Simple main effects were used to establish if any feature was affected differently by its replacement number, and if any replacement was affected differently by which feature was being replaced at that point. These were calculated for each replacement (1-6) with the between-subjects factor of system (Picasa, Lightroom, Twinsornot). Results showed no effect of system on replacements 1, 3, 5 and 6 ($p > .052$) but a significant effect on replacements 2 ($p = .043$) and 4 ($p = .035$): simple contrasts showed this effect could be attributed to Picasa yielding significantly higher miss rates than the other two systems ($p < .031$). For replacement 4, Twinsornot yielded significantly higher miss rates than Lightroom ($p = .010$).

Second analysis: As before, the analysis was repeated using three other veridical images for only Picasa and Lightroom.

Table 6.3-8: Phase 2 – Mean miss in percentage by replacement (comparison of first and second analysis).

		Picasa		Lightroom	
		1st	2nd	1st	2nd
Replacement	1	4.2	70.8	0.0	20.8
	2	12.5	58.3	0.0	25.0
	3	8.3	66.7	0.0	29.2
	4	8.3	79.2	0.0	50.0
	5	25.0	91.7	25.0	70.8
	6	70.8	100.0	75.0	87.5
	Mean	21.5	70.8	16.7	47.9

Mean miss scores (%) for Picasa and Adobe Lightroom. n=24 per system. Column '1st' lists the means for the first analysis using the original veridical target images. Column '2nd' lists the means for the second analysis using three new veridical target images.

Descriptive analysis: Results, shown in Table 6.3-8, show an incline in miss rates for both systems between the first analysis (1 original image) and the second analysis (3 new images). For the Picasa system, a mostly ordinal increase in miss rates can be seen for the second analysis, except for replacement 1. The first analysis showed a similar pattern, except for replacement 2. Lightroom yielded floor results for replacements 1-4 followed by ordinal increases for replacements 5 and 6 in the second analyses, but an ordinal increase for the first. For replacement 6, Picasa yielded no correct recognitions, but Lightroom yielded three (accounting for 12.5% of items).

Analysis of Variance: Each system was analysed separately using Univariate ANOVA for the factor of replacement (0-7) and the second analysis data (3 new). Picasa showed a significant main effect of replacement (*F(*7, 184) = 21.29, *p* < .001, η_p^2 = .45). Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into four subsets (see Table 6.3-9)

Table 6.3-9: Phase 2 celebrity - Homogenous subsets for replacement using three veridical images (Picasa)

Replacement	N	Subset			
		1	2	3	4
0	24	0.0			
2	24		58.3		
3	24		66.7	66.7	
1	24		70.8	70.8	70.8
4	24		79.2	79.2	79.2
5	24			91.7	91.7
6	24				100.0
7	24				100.0
Sig.		1.00	.42	.20	.07

A significant difference occurs after replacing at two features, with the miss rates non-ordinal. A further significant increase occurs after replacement 3 and 1 (moved to 4th in the order).

Lightroom also yielded a significant effect of replacement (*F(*7, 184) = 920.36, *p* < .001, η_p^2 = .44). P Tukey's HSD Homogenous subsets for the factor of replacement shows data grouped into four subsets (see Table 6.3-10).

Table 6.3-10: Phase 2 - Homogenous subsets for replacement using three veridical images (Lightroom)

Replacement	N	Subset			
		1	2	3	4
0	24	0.0			
1	24	20.8	20.8		
2	24	25.0	25.0		
3	24	29.2	29.2		
4	24		50.0	50.0	
5	24			70.8	70.8
6	24				87.5
7	24				100.0
Sig.		.15	.15	.56	.15

A significant difference occurs after replacing at least four features with a second significant change occurring after replacing five, followed by replacements 6 and 7, which are grouped together.

Comparison of analyses: For the Picasa system, Univariate ANOVA for the factor of replacement and method of analysis (1st, 2nd). A main effect method of analysis was found to be significant (*F*(1,368) = 45.00, *p* < .001, η_p^2 = .46). This may have been due to amplitude differences in miss rates: second analysis was higher than the first.

An approaching significant interaction was found between replacement and method of analysis (*F*(7,368) = 9.84, *p* < .001, η_p^2 = .16). Previous analyses showed an effect of replacement for both the first and second methods. The effect of method of analysis was analysed for each replacement separately. All replacements showed an effect (*p* < .005): the second analysis yielded higher miss rates than the first.

For Lightroom, the ANOVA was repeated and found a main effect of method of analysis to be significant (*F*(1,368) = 65.65, *p* < .001, η_p^2 = .56): the second analysis yielded higher miss rates than the first.

A significant interaction between replacement and method of analysis was found ($F(7,368) = 4.32$, $p < .001$, $\eta_p^2 = .08$). The effect of method of analysis was analysed for each replacement separately. Replacements 1-5 showed an effect ($p < .019$) where the second analysis yielded significant higher miss rates than the first, but replacement 6 did not ($p = .28$).

Comparison to humans: As before, the automated systems second analyses data was compared to the human participant data (celebrity).

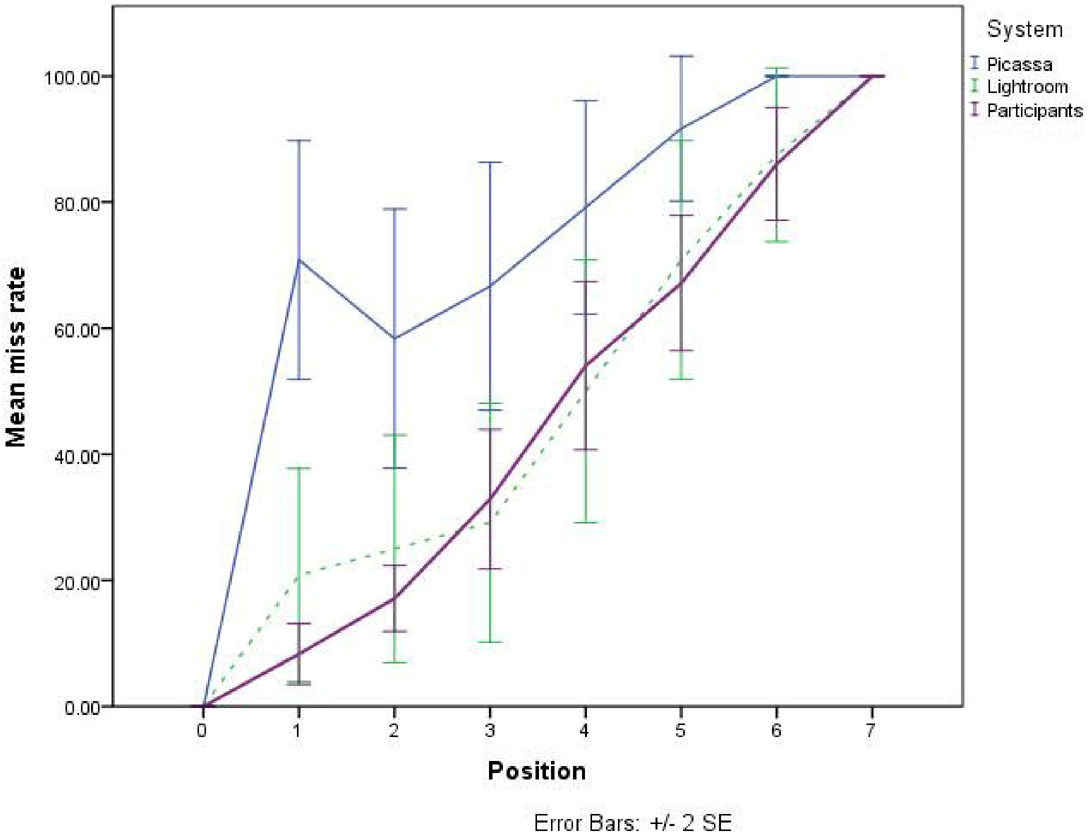


Figure 6.3-2: Phase 2 (participants and automated systems) - Mean miss in percentage by replacement

Descriptive analysis: Figure 6.3-2 shows that overall mean miss rates were lowest for the participant data (*M* = 44.3%) similar to Lightroom (*M* = 47.2%) compared to Picasa (Picasa, *M* = 77.8%) which was considerably higher. The ordinal increase in miss rates as features are replaced is observed in the Lightroom and participant data. This is also mostly the case for the Picasa data except for replacement 1 (*M* = 70.8%) is higher than means for replacements 2 (*M* = 58.3%) and 3 (*M* = 66.7%). To note, Picasa also yielded

100% mean miss rates for when all six features had been replaced, in comparison to the participant data (*M* = 86.1%) and Lightroom (*M* = 87.5%).

Analysis of Variance: Univariate ANOVA for the between-subjects factor of replacement and factor of system (humans, Picasa, Lightroom) found a significant main effect of replacement (*F*(7,552) = 79.85, $p < .001$, $\eta_p^2 = .50$) and system (*F*(2,552) = 36.35, $p < .001$, $\eta_p^2 = .12$). Post-hoc Tukey's HSD analysis of systems found Picasa yielded significantly higher miss rates than both participants and Lightroom ($p < .001$). However, no significant difference was found between participants and Lightroom ($p = .78$).

Simple main effects: A significant interaction between replacement and system was found (*F*(14,552) = 3.13, $p < .001$, $\eta_p^2 = .07$). Simple main effects of system for each replacement were carried out. All replacements showed a significant effect of system ($p < .043$) and an approaching significant effect for replacement 6 ($p = .081$): replacements 1-3 revealed that Picasa showed significantly higher miss rates than both other systems ($p < .019$); for replacement 4, Picasa showed an approaching significantly higher miss rate than Lightroom ($p = .052$); for replacement 5, Picasa was significantly higher than participants ($p = .045$); and for replacement 6, no system was significantly different to the other ($p > .10$).

6.4 Discussion

For the purposes of this chapter, the results from the automated face recognition study will be discussed separately for phase 1 and 2. It was not known whether performance for the automated systems or human participants would be better due to them both potentially having advantages outlined in the literature: Research had found that performance for frontal face images was superior for the automated systems, however other research found that humans were superior when using familiar faces when parts had been disguised (see section 2.6 Automated Face recognition systems for a review). Two main themes emerged from phase 1:

1. The automated systems were heavily reliant on the images used for training,

2. Each system was differently affected by featural manipulations which were also different from the human participant data.

Three main results emerged from phase 2:

1. The automated systems were reliant on image training,
2. Systems showed a mostly ordinal increase in miss rates through feature replacement and although different in score amplitude, were comparable to the human participant data,
3. One system yielded correct recognitions when all six features had been replaced.

Phase 1: Results showed that for phase 1, Lightroom and Twinsornot systems were not affected at all by the featural manipulations, yielding 0.0% miss rates/dissimilarity across all conditions. Picasa, did however, yield some misses with an overall miss rate of 38.2%. This analysis was carried out by training the systems on the original veridical image that the condition stimuli were generated from. Because of this, it is likely that the systems used image pattern recognition rather than face recognition, so it is unsurprising that the miss rates/dissimilarity scores were so low. Therefore, the analysis was repeated using three additional images of the celebrity targets for system training for Picasa and Lightroom. Overall miss rates were higher for the second analysis, and this reflects a more real-world scenario, where faces are familiarised with several views and contexts. Picasa showed the largest increase in miss rates (+22.1%) compared to Lightroom (+16.7%).

In the first analysis, Lightroom and Twinsornot yielded floor-level results and so no feature hierarchy was present. For Picasa, however, there was a feature hierarchy with a significant effect of feature on miss rates. Replacing the mouth yielded the lowest miss rates and therefore could be considered the least salient for Picasa. This was followed by the outline, nose, and eyebrows, with the hair and eyes showing the highest miss rates and therefore highest saliency for the system. Post-hoc tests revealed two notable results: significant differences between the most salient eyes and all other features, and the hair and all other features, except the eyebrows, nose and outline. The least salient feature, mouth, was also significantly different from all features, except the nose and outline. This suggests a grouping of the highest saliency features (eyes and hair) compared to the rest, and the mouth in its own low importance group.

In contrast, the second analysis yielded a hierarchy for Lightroom and a different one for Picasa (Twinsornot not analysed), in addition to higher miss rates. Lightroom now showed hierarchies for the features, but no significant effect of feature, and the hierarchy for Picasa (showing a significant effect of feature), differed from the first analysis: with the eyes dropping from the highest saliency to the lowest for the second analysis. Lightroom's hierarchy showed two notable differences to Picasa: the eyes were the most salient and the hair now the least. The lack of effect of feature for Lightroom suggests that even though a hierarchy was present, the differences between features were not vast. This suggests that the Lightroom system is perhaps not overly reliant on feature processing but rather uses a more holistic (overall image) approach.

In comparison, participant data showed a lower level of miss rates compared to Picasa and Lightroom (using second analysis data, excluding Twinsornot), suggesting that participants perform better and are more invariant to featural manipulations. However, participant (9.1%) and Lightroom (16.7%) overall mean miss scores were not statistically different. Picasa (60.3%) was, however, significantly higher than both participants and Lightroom with majority miss rates. This suggests that the holistic processing strategy used by humans for familiar faces yielded an advantage for human performance in comparison to the automated systems. Again, hierarchies varied with notable differences as the least salient feature for participants, eyebrows, shifted to second highest for Picasa and Lightroom. Eyes were consistently important for both participants and Lightroom but low importance for Picasa, demonstrating vast differences in each systems reliance on features. In support of this, Picasa showed statistically significant differences for all features when compared to the other systems (except for the eyes).

Results suggest that both systems are highly dependent on the images with which they are trained, indicated by the statistically significant differences found between the two analyses for both Picasa and Lightroom, driven by differences in score amplitude. There were also differences in the saliency of featural information between systems. Overall miss rates were moderately high suggesting that should these types of composite images be run through these systems, there is a reasonably high likelihood that they will not be recognized. Additionally, this poor performance is likely because the systems are reliant on the use of veridical target images in training. Results also suggest that changing one feature is enough to yield a similarity score that is below the predefined threshold of what is classified as a ‘match’ by the systems.

Phase 2: Overall performance for phase 2 was fairly good with low miss/dissimilarity scores for all systems (Picasa, 21.5%; Lightroom, 16.7%; Twinsornot, 16.0%). For all systems, replacement 1 always yielded the lowest score and replacement 6 the highest, as expected, with a significant effect of replacement for all systems (all systems showed at least linear and quadratic patterns). However not all increases were completely ordinal with replacement 2 lower than replacement 3 and 4 for the Picasa system, although a linear/quadratic/cubic distribution pattern was found. All systems also showed statistically significant differences between replacement 6 and all other replacements with Lightroom and Twinsornot yielding more widely spread feature scores with significant differences between replacements 5, and sometimes 4, against all others. Feature replacement was clustered in significantly different groups differently for each system, suggesting that the potential identity criterion points shifted between systems. Despite these differences, the systems were not statistically different from one another. Only two replacements (2, 4) showed an effect of the system being used (replacement 2, between Picasa and both other systems; replacement 4, between Lightroom and Twinsornot).

As for phase 1, the analysis was repeated using three additional veridical images for training with Lightroom and Picasa. Overall miss rates increased for both systems with Picasa yielding the largest increase (+49.3%) compared to Lightroom (+31.2%), supported by a statistically significant effect of the type of analysis. Both systems showed a mostly ordinal increase in miss rates during feature replacement, except for Picasa, where replacement 1 yielded lower scores than replacements 2 and 3. The effect of replacement was, as expected, statistically significant and linear in distribution and suggests that the systems were affected by the quantity of veridical detail left in the face image. Lightroom's data was more widely spread (more statistically significant differences between features) suggesting a stronger effect of feature replacement. Similar to the participant results, miss rates were not at ceiling level once all features had been replaced at replacement 6 for the Lightroom system (3 items), suggesting that some residual information was providing identity cues for recognition (see Figure 6.4-1).

Figure 6.4-1: Targets receiving correct recognition for replacement 6 (Lightroom) using three other veridical images.

Three targets were correctly recognised by Lightroom when shown a target with all six features replaced. (Left: three other veridical images used for training; right: replacement 6 composite).

It is possible that the configural and contrast information leftover in the stimulus was sufficient for this, without any featural detail. Looking at the replacement 6 stimuli, there are strong resemblances to the target due to remaining configural and contrast information, but also some remaining face type information (e.g. Hugh grant has a long narrow face and a long chin, Leonardo DiCaprio has a widow's peak).

In comparison, participant data (42.3%) showed a lower level of miss rates to Picasa (70.8%) and Lightroom (47.9%) (using second analysis data, excluding Twinsornot), suggesting participants perform better and are more invariant to featural manipulations. This overall effect of system was significant and, additionally, a significant interaction identified Picasa as yielding significantly higher miss rates than both other systems, but not between participants and Lightroom. All systems showed an ordinal increase in miss rates as expected (except Picasa where replacement 1 yielded lower miss rates than replacements 2 and 3). It seems that in terms of processing, Lightroom was most similar to the participant data, with both showing an entirely ordinal increase in miss rates during feature replacement and continued recognition of identities even when all features had been replaced at replacement 6, suggesting both were using other streams of information: probably configural and contrast. They both also suggest that a theoretical identity criterion point occurs once three (participants) or four (Lightroom) features have been replaced (half the face or more).

To summarise, the automated systems performed better than participants (lower miss rates) in both phases when they were trained on the original veridical image that all stimuli were generated from. This suggests that the systems were using pattern rather than, or in addition to, face recognition type processing. When the automated systems were trained on three additional images of the celebrities instead, miss rates increased significantly, and performance dropped to below that of participants as indicated by overall mean miss rates. Phase 1 yielded feature hierarchies for Picasa and Lightroom (2nd analysis), although only significant for Picasa. Lightroom's hierarchy was most comparable to participants with the eyes considered most salient. In phase 2, Lightroom was more comparable to participants (celebrity data) with an ordinal increase in miss rates, a first criterion point around the middle of feature replacement and correct recognition for when all six features had been replaced.

7 General Discussion

Created faces provide a way to generate new identities, or adapt existing ones, for various applications such as for animation, CGI, training purposes, and stimulus sets to name a few. The emergence of digital applications and technology has brought about a way to digitally generate these types of faces, either through manually creating a new face, such as in CGI scenarios, or through manipulating existing face images. This study focused on scenarios where face features are sampled from donor faces (photographs) and composited to form a new one. Two applications of this were used as motivation for the research:

1. Creating new faces that behave as avatars on fake social media profiles for the covert surveillance of online criminal activity.
2. Sampling from face photographs to 'texture' forensic facial depictions.

From these applications, the question arose as to whether compositing conceals the identity of the donor component parts. This becomes important when sampling from face photographs as the donors may not wish to be identified, especially in the forensic applications mentioned above. The study used familiar face recognition tasks to test stimuli where familiar faces had had facial features replaced with unknown ones, to test whether replacing some features concealed identity more than others (Phase 1). A second experiment (Phase 2) replaced features in a compound manner to see how many features needed to be replaced in order to conceal identity.

It was hypothesised that:

1. a facial feature saliency hierarchy would be found when replacing target features individually in a whole face.
2. replacing features in a target face in a compound manner would yield an ordinal increase in miss rates and that there may be a criterion point during compound feature replacing where the identity shifts from old to new.

3. some features may be more important to be replaced and may affect criterion points.

Additionally, more memorable faces may be more invariant to the featural manipulations of the compositing process (see section 2.7 Summary for the full hypotheses).

Phase 1 was split into two parts (phase 1i and 1ii). Phase 1i showed whole faces with features replaced and phase 1ii showed isolated target features to assess the relative importance of each feature in embedded and isolated conditions. To summarise the results, phase 1i yielded a feature saliency hierarchy, revealing eyes to be the most impactful when changed, and eyebrows the least. Celebrities and lecturers represented different familiarity types as indicated by the external feature preference for lecturer targets (lower familiarity) compared to the internal preference for celebrities (higher familiarity). Combined data for phase 2 revealed a completely ordinal increase in miss rates by feature replacement with a theoretical criterion point from old to new identity indicated by a significant increase in miss rates. Again, familiarity affected miss rates, with celebrities yielding overall higher miss rates due to the more impactful disruption of the compound feature replacement to holistic processing. The reverse of this effect was found in phase 1i (lecturer miss rates higher than celebrities) where single feature replacement did not disrupt holistic processing to the same extent. Overall, the most salient feature from phase 1i (eyes) was also found to be more important in phase 2 where it yielded significantly higher miss rates when replaced, compared to the mouth. The effect of feature was also found to be more impactful in the middle of feature replacement where miss rates increased significantly. Identity was not concealed all of the time indicated by correct recognitions for when all features had been replaced, suggesting other types of residual information were being used.

This chapter will discuss overarching themes across the experiments, followed by limitations of the study, contributions to theory and practice, and potential future research. The following themes emerged from the results:

1. Configural information – featural information was manipulated, but configural detail may have facilitated recognition.

2. Contrast – Residual contrast information may have provided cues to identity
3. Stimulus Specific results – Interactions between some characteristics and miss rates were observed. Are target faces impacted differently by featural manipulations depending on their characteristics?
4. Explicit choice – Overt type recognition results mostly supported the recall results

7.1 Stimulus specific results

This section discusses any stimulus specific effects (characteristics or target sex appearance) across all phases of the experiments.

Characteristics

Each of the celebrity targets was rated for three different characteristics that could be used as a covariate in the experimental analysis to look for specific stimulus effects. A subsequent familiarity Control section was also included to assess if there was evidence for whether familiarity with the target affected participants' perceptions of these characteristics. The three characteristics were rated on a 7-point Likert scale so that an overall (Mean) assessment of rating could be found for each target. Increased familiarity resulted in higher ratings for memorability.

Phase 1i showed a significant effect of memorability: the higher the memorability score the lower the miss rates overall, as indicated by a scatterplot and a significant negative correlation. It seems that the more memorable or distinctive a face is, the more invariant it is to changes made to individual features. Figure 7.1-1 shows the top and bottom three scoring celebrity targets for memorability.

Highest memorability scores

Angelina Jolie Leonardo DiCaprio Hugh Grant

Lowest memorability scores

Benedict Cumberbatch Hugh Jackman Tom Hardy

Figure 7.1-1: Memorability of celebrity targets.

Top row: the top three scoring celebrity targets for memorability. Bottom row: the bottom three scoring celebrity targets for memorability.

If a target face is more memorable, this may render the face less vulnerable to featural-type changes. Faces were rated as a whole, and therefore there is no way to know what is driving the memorability of the faces, and if that is feature specific. For example, Angelina Jolie is known to be memorable because of her lips. It is also possible that the memorability for any face may not be driven by features but could be driven by more holistic impressions such as a memorable configuration of features or an interesting level of contrast across the face (Logan et al., 2017, Valentine, 1991, Valentine and Bruce, 1986). An approaching significant interaction between memorability and feature was found, suggesting that featural manipulations are affected differently by this characteristic. Main effects showed an increased effect of memorability for (higher memorability, lower miss rates) when the facial outline was replaced, compared to other features. Therefore, it seems that if the remaining facial information is highly memorable, then changing the outline feature will have little effect, suggesting a low level of saliency for the outline and that perhaps this feature is not as important for perceptual measures of memorability. Approaching significant effects of memorability for eyes and hair was found. Although not actually significant, this is a more expected result as hairstyles may be memorable and particular to a person's identity. In contrast, hairstyles are often changed, in particular for celebrities, and so it could also be argued that this result is surprising (Wright and Sladden, 2003).

The isolated feature version (phase 1ii) also showed a significant negative effect of memorability overall where the more memorable the face is, the lower the miss rates (indicated by a scatterplot and significant negative correlation). It can, therefore, be theorised that the effect of memorability is not just interpreted from the whole face, but also continues through to individual features and the lack of a significant interaction shows that this effect was relatively even across feature conditions. It seems that the encoding of faces, in general, allows for different levels of memorability that facilitates recognition of features presented out of the whole-face context.

Unlike phase 1, phase 2 showed no effect of memorability. This could be explained by the compound feature replacement effect cancelling out any memorable residual information that may have been used. Phase 1 only replaced a single feature, and so it is possible that there was still enough information remaining in the face that facilitated recognition for high memorability faces, but less so for low memorability faces. This was the case for the early replacement conditions (e.g. replacement 1) where a considerable amount of the target face was still remaining but any effect was reduced with the increase in feature replacement. In contrast, phase 2 did find an overall effect of trustworthiness: the higher the trustworthy score, the lower the miss rates. However, no correlation was found between the mean miss of targets and their trustworthiness score. In contrast, trustworthiness and attractiveness were not found to have an effect on miss rates for phase 1i, nor was attractiveness for phase 2. It seems that these characteristics, even though found to be embedded in recognition of faces in previous literature (Botwin et al., 1997, Little et al., 2006), are less useful under these types of manipulations, in the context of this study.

Target faces were rated for overall memorability which should have, in theory, controlled for any distinctive features that may affect individual feature results. However, it is possible that one distinctive feature does not render a face as memorable, in general (Benson and Perrett, 1994, Bruce et al., 1994). It is possible that requesting memorability scores for each feature may have provided more insight into the role of salient features and the effect of distinctiveness. However, this would have been an extensive task to setup and outside the scope of this research.

Target sex appearance

Target sex appearance, in general, did not have an effect on the results across experiments. However, in the isolated feature experiment (phase 1ii), an interaction between target sex and feature was revealed. This was attributed to the hair condition with males yielding significantly higher miss rates (65.8%) than females (48.1%) demonstrating a higher saliency for female hair than male, in terms of identity, when presented in isolation. This effect was not found for the whole face version of phase 1i both overall and as an interaction with feature. This suggests that the relative saliency of hair for each sex only diverges when the feature is out of a whole face context. It is possible that in a whole face context other cues are used for determining the identity of the individual due to a robust memory for that face where the internal features are utilised more than the external ones. However, if other feature cues are not available then hair becomes more important, but only for females. Hair also represents a feature of great variation, particularly for females where hairstyles can be modified on a daily basis. Because of this, it could hypothesised that hair would be too variable for identity recall in isolated conditions, so this result is somewhat surprising in that respect. No other phases showed an effect of target sex; however, phase 2 lecturer data did not have enough responses for male targets to perform inferential statistics between males and females.

7.2 Explicit choice

An Explicit choice task was included after the Spontaneous naming task to allow for an overt type recognition to take place to compare with the recall type recognition results. This was not designed as a separate type of study, but to allow participants to respond by giving multiple choices from which to choose from, to mimic a more real-world type scenario and to prevent ‘familiar’ responses from being placed into the ‘unfamiliar’ response option. Collecting overt explicit choice scores also allowed for a comparison between the spontaneous naming miss rates and overt scores. The celebrity version of phase 1i and both versions of phase 2 found significant positive correlations between spontaneous naming miss rates and explicit choice miss rates, as expected: the higher the spontaneous naming miss rate, the higher the explicit choice miss rate. If a participant can correctly recognise/recall a target in the Spontaneous naming task, he or she should be able to correctly choose the target name in the Explicit choice task, and vice versa for miss rates. It was unknown whether a participant who could not recall identity specific information, would not be able to pick out the correct target name in the Explicit choice task due to unfamiliarity or just because they could not recall identity specific information in the Spontaneous naming task. The correlation suggests the former. This was not the case for the isolated version of phase 1 (1ii) where no reliable correlation was found. This would suggest that because the isolated feature task was so difficult, name cues in the Explicit choice task may have facilitated correct overt recognition to a greater effect than for the whole-face versions. Participants in the LJMU group for phase 1, always correctly answered the Explicit choice task for all valid trials, and therefore there was no variation across items. Analysis was carried out on only the UCLan data which showed no reliable correlation, however, items were low (n=6).

The Explicit choice task did, however, appear to incur some bias in the design. Only two distractor names accompanied the target name. Therefore, it is possible that participants used an elimination process through familiarity with the distractor names to allow isolation of the correct target name. Therefore, any correct recognition results from the Explicit choice task could have been inflated because of this. It is also possible that prior exposure of the face during the Spontaneous naming task may have provided some priming as to who the face might belong to, making the Explicit choice task easier than if it had been shown in isolation. Distractor names were not assessed for their similarity to the target with respect to the visual appearance of the face, nor for the similarity in the name itself. However, distractor names were only presented once per target and randomised in presentation across the stimulus set to balance out any name similarity effects that may have occurred. The names were matched for target sex appearance which reduced the pool from which participants needed to extract a face memory.

7.3 Streams of information

The experiment was designed so that the miss rates could be attributed to the featural change only as all stimuli were generated with the original target configuration and overall contrast pattern. This was to try and isolate the featural information being changed so that one stream of information was being tested in the study (see section 2.7 Facial creation for a summary of this). However, it is likely that these other streams of information were inadvertently altered, and may have facilitated recognition (Burton et al., 2015, Cabeza and Kato, 2000) (see section 2.4 Holistic processing for a review of the different streams of information used for face recognition).

Configuration

Care was taken to ensure that the overall position of the features remained consistent with the targets, so that configural information was controlled for. However, it is inevitable that changing featural information will inadvertently alter the configural or interdependent relationship of the features. For example, when the eyes were changed, the pupil position was matched exactly with the replacement feature and the general fissure width also kept constant. However, if the target had a fairly 'closed' fissure shape where the height between the lower and upper eyelids was small and this was replaced with eyes with an 'open' fissure shape (larger distance) this will of course have changed the relationships and distances on the vertical plane between the eyes and other features. Another example may involve the mouth (lips). These were matched for commissure position and width, however the thickness of the lips would not have been controlled for and, again, may have altered the interdependent relationships between the lips and other features on the vertical plane. Note that, all features were matched for their position on the vertical plane as well as their dimensions on the horizontal plane only: their dimensions on the vertical plane were not matched (except the nose) which may have been reflected in the low miss rates. Although features' positions were the same on the vertical plane, their relative distances were affected by featural dimension changes.

Contrast

Image contrast properties across the face was maintained between the target face and the unique composite to isolate testing to only the featural stream (see section 2.4 Facial contrast and pigmentation for a review on facial contrast). As configuration was also maintained, the contrast pattern across the face should have also remained constant. It is possible that small featural changes will have affected shape from shading information (which makes up some of the contrast information) within that feature that may have removed contrast recognition cues. Features were not matched for any kind of general shape, only general contrast, and so it is also possible that a deep-set eye may have been replaced with a more prominent eye, thus changing the shape from shading information and effective contrast of that feature and its surrounding contour. Therefore, although care was taken to minimise changes to other streams of information useful for face recognition, altering featural information will still inadvertently affect these in small amounts. The results from phase 1 suggest that changing one feature does not affect recognition drastically due to minimised disruption to the rest of the face and any inadvertent changes may also not affect this. However, it is possible that the miss rates would have been even lower had features been matched for similar shape and structure as well, which would have meant very little featural change. Swapping features but matching for contrast, configuration, general shape and structure is also impractical and ecologically redundant due to the amount of care needed to be taken in a real-life scenario. However, it is possible that any practitioner using compositing for concealing identity may be more inclined to choose replacement features that are similar to the targets, thus minimising the effect compositing may have on concealing identity. This is something practitioners should probably bear in mind when generating composites.

7.4 Limitations

The study design and methodology may have resulted in limitations that impacted on the results, and these will be discussed here.

Compositing technique

The compositing technique varied in its success in creating plausible faces that appeared unedited. Restrictions on the time able to be spent on each composite stimulus was short due to the large number of stimuli to be generated (over 450 composite stimuli) in a relatively short period of time, meaning that the necessary image manipulation and refinement needed to create a plausible unedited face was not always possible. The ramifications of this is that participants may have detected the edited portions of the images and subsequently ignored them in their processing, to allow for recognition of the other parts of the face. Given more time per composite stimulus, this effect may potentially have been reduced.

The compositing technique aimed to only manipulate featural information whilst maintaining the configuration and contrast of the target. Because of this, it is possible that when features were composited onto the target face, too much care was taken to control for these variables that has resulted in replacement features appearing very similar to the target. It is also possible that there is some kind of innate bias to composite features to appear similar to the target face so that the face appears plausible. This may have rendered the stimulus exemplars more similar to the target face, making the task easier. This, in turn, could have facilitated the correct recognition of stimuli where all six features were replaced. However, the positive advantage of this technique meant that composite stimuli were plausible.

Sample size

The sample size of target faces for the celebrity version was considerably larger than that for the lecturer version. There were limitations in lecturer target recruitment due to the difficulties in recruiting persons willing to donate their face images and donating the time to be photographed and consulted for further information. The requirement for the lecturers to be familiar to large group of students also made recruitment more restrictive. Similarly, recruitment of participants that would be familiar with the lecturers was restrictive: Familiarisation needed to occur over at least one academic semester so that the participants were more likely to have a familiar memory of the target lecturers. Because of this, participant recruitment and testing was restricted to mainly the second semester and some of the third semester for post-graduate students. Data collection was reliant on the good will of participants who were interested in participating as no remuneration was provided. However, for the UCLan participants, course credit was provided for participation with a requirement for participation within the programme of study. Therefore, recruitment for the UCLan participants/lecturers was less restrictive than for the LJMU group.

As a further note, the overall mean miss rate for celebrity data was higher than lecturers for phase 2, but less for phase 1. It is possible that the low numbers of participant recruitment for phase 2 (n=71), has driven this result, compared to the higher numbers for phase 1 (n=148). Low power for the lecturer data may have affected the clustering of replacements into the homogenous subsets. However, effect sizes (Cohen's d), that are independent of sample sizes, were calculated by position for celebrities and lecturers and showed differences between the two types, with celebrities yielding a higher mean effect size than lecturers.

Participants

There were some limitations with the participant pool recruitment, listed as follows:

Age: The celebrity targets were selected through a pilot study requesting the most popular celebrities. The demographic used to recruit from for the pilot study was the same as that used for the experiments: mainly students. Because of this, the participant group mean age across experiments was relatively young (<30 yrs.). This meant that the participant pool was, potentially, not a fair representation of the broader age range in the general public. However, it was important to recruit participants who were most likely to be familiar with the target face selection as indicated by the pilot study. Should a broader age-range of participants been recruited for the pilot study, it may have been difficult to select a popular subset with which to test and any subsequent broad participant pool may have been more varied in their likelihood to be familiar with the target face selection. The positive angle to this was potentially higher and more consistent familiarity with the target identities.

Gender: Due to the demographics of the participant pool, 76.3% of the experimental participants were female. This is perhaps not a true representation of the more even share of males and females in the general population. Research has also shown that women are better at recognising female faces, (Loven et al., 2011, Herlitz et al., 2013, de Frias et al., 2006), therefore one might have expected lower miss rates for female target faces. However, the results showed that target sex appearance did not affect results (only in the isolated hair condition).

Other-face effects: All participants were invited to participate in the studies, without excluding individuals of a certain age, sex appearance or ethnicity. As outlined in section 2.5 Other-face effects, there are some qualitative differences between the processing of faces that we are less likely to have encountered on a regular basis during our lifetime. Valentine’s face space model (see section 2.5 Face memory and storage) suggests that these effects occur because the faces sit further outside of our own-face space and therefore we find them more difficult to recognise. Research has also shown that we may adopt different face processing strategies for face types that we do not normally encounter (e.g. cross-race effect). For example, during target face selection, the faces were narrowed down to white European faces as this was the predominant ethnic group in the demographic population being tested. This was to avoid the inclusion of highly distinctive faces that sit outside of the participant’s own-face space. An assumption was made that the population was made up of individuals that had been residing in the demographic population and would therefore have a level of familiarity with the target face population, even if the participant had normally been residing in another ethnic group population. However, it is worth noting that even if the target was familiar to the participant, their general own-face space may not have normally encountered that type of target face and therefore noise may have been added when that participant was shown a composited version of the target, due to using a different processing strategy. Therefore, it is entirely possible that noise has been added to the data, due to adoption of different processing strategies for not normally encountered face types.

Face recognition ability: Differences in face recognition ability could account for some differences in miss rates occurring through the replacement of a single feature. Participant ability at face recognition was not tested prior to or after the experiment. It was important to collect a data set across a population of individuals with a wide variety of abilities (see Appendix A - Methodological considerations, for more detail). This is more in line with a real world scenario where the choice of who sees the avatar is uncontrolled, although this may have caused some noise in the data. For example, excluding prosopagnosic individuals, who take up 2% of the population (Jiang et al., 2011), will have reduced noise but would not be a realistic representation. It has also been shown that prosopagnosics may adopt a more piecemeal approach to face recognition and as such may have found the task slightly easier by being able to exclude incorrect featural information that normal recognisers may not be able to due to the overwhelming holistic processing technique. In contrast, any super recognisers may or may not have found the task difficult due to their proficiency in face recognition. However, it is also possible that their known expertise in holistic processing may have rendered any manipulated face as a new identity. Subsequent analysis for participant ability was also not carried out as the information would not have resulted in any participants being excluded based on either really poor or really good face recognition ability. Including all trials does of course result in the risk of experimental noise through the large differences in face recognition ability and the possibility that individuals were adopting different qualitative face recognition strategies depending on their ability (Davis et al., 2016, Wang et al., 2012, Wilmer et al., 2010). It seems that the decision to not test participants face recognition ability, prior or after, the experiment may have, in fact, impacted on miss rates results. This could be evident in the surprising result where some participants were still able to recognise the face, even after all six features had been replaced. As discussed above, this may be due to participants using residual information left in the image, or it could be due to those participants having superior face recognition abilities and therefore being more invariant to featural manipulations. In a real world scenario, it may be that it members of the public with superior face recognition abilities are the very group that may be able to identify donor features in created faces, and therefore having data to group the participants in this way, would have helped to establish whether this was the case. Furthermore, not having participants face recognition ability data, meant that participants could not be compared to one another in relation to their performance. For example, in the Automated face recognition study, the three systems were able to compared to one another. However, without data that could separate the participants into different ability groups, comparisons could not be made in the same way for humans. It is additionally worth noting this limitation, so that should further research be carried out, testing participants' face recognition ability should be implemented into the experimental design.

Online testing

Online testing was used to reach a large audience to generate the high participant numbers needed for the studies. Research has shown that results are not significantly affected by the mode in which the participant data is collected (Jones et al., 2007, Metzger et al., 2003). However, it would be advantageous to carry out the whole experiment under laboratory conditions to control for various extraneous variables, including viewing distance (others may include lighting, noise in the room, screen size etc.). For the current study, this lack of control for viewing distance did not outweigh the advantage of recruiting from a larger participant pool to generate higher numbers of participants (see Appendix A - Methodological considerations for more detail on the use of online testing methods in face perception research).

Qualtrics sampled a condition item from the six available for each item and this sampling was random. Therefore, participants saw different numbers of condition stimuli across the experiment, but always only one stimulus per item. This, coupled with the extraction of only familiar (valid) trials, resulted in uneven sample sizes. This limitation of the design could be resolved for future research where further exploration of Qualtrics' capabilities would allow for control over how items were seen. However, this is a small issue as ANOVA has been shown to be relatively insensitive to difference in cell numbers. This analysis was carried out by items and so differences in the number of observations per cell is likely to be low (except lecturer data for some cases). For future reference, it is possible that another experimental platform could be used for data collection, or, the design revised in way that would facilitate support from the Qualtrics system.

Experimental design

The experiments were setup to allow for different types of familiarity to be recorded. An unfamiliar response was used to indicate that participants did not recognise the stimulus. Familiarity and recall was used to allow participants to provide a text input to prove that they were familiar with the target and could provide evidence of recalling the correct identity. This response was classed as a correct recognition (if the correct name/description was given, if incorrect it was marked as a miss as the face was clearly reminding them of someone else) and the former was classed as a miss and it was the miss data that the analysis was carried out on. However, an intermediate response of 'familiar' was implemented to allow for participants to indicate that they felt that there was something about the face that reminded them of someone but could not quite place the identity. These responses are less important to the question of whether identity was concealed in that miss responses are a more stringent version of 'unfamiliarity' than a 'familiar' response is (Bruce and Young, 1986). However, the 'familiar' responses do provide an indication of some kind of memory for a face, but without semantic recall there is no way of knowing if the stimulus is reminding them of the correct identity or someone completely different. However, it was important to separate these types of responses from the 'unfamiliar' ones so that the miss rates were not over-inflated. Therefore, any miss rate results could be generalised to a real-world setting in a more generous way.

It should also be noted that the Ratings study that was used to assess a target faces' characteristics as a way of explaining any stimulus result, was based on a holistic impression of the whole face. However, the compositing manipulations involved replacing individual features. Therefore, it is possible that miss rates would have differed between the replacement of more or less characteristic features, even if the target face had been given a different overall characteristic rating. This limitation may have reduced the impact of the holistic characteristic ratings scores when analysing the miss rates. However, the ratings are still valuable in that they gave some insight as to whether holistically characteristic faces are more or less invariant to featural manipulations. As mentioned in section 2.5 Face memory and storage, one study found no differences in recognition rates for featural manipulations across different face distinctiveness levels. However this was an unfamiliar discrimination task and the results may not be generalizable to a familiar scenario. In light of this, it seems that further research, collecting ratings on individual features, may help to explain the impact of replacing more or less characteristic individual features.

7.5 Future research

Based on some of the limitations of the current study and some of the surprising results yielded from it, four potential future experiments are proposed:

1. It would be interesting to investigate the characteristics of each feature so that they could be directly compared to each feature replacement. Although a laborious task, the experiment may ask participants to rate the characteristics of each feature rather than the whole face. For example, a face with very memorable lips, such as Angelina Jolie, may yield a higher miss rate when that feature is replaced. However, there is always the possibility that a face's characteristic is based on more holistic opinions (Cabeza and Kato, 2000, Schwaninger et al., 2002, Tanaka and Sengco, 1997). For example, George Clooney's facial morphology is not overly memorable, but his face is. What seems to be driving this memorability is the pigmentation and contrast pattern of his face. As another example, what makes Bruce Willis' face memorable is the large distance between his nose and mouth (configural). Therefore the study could be extended to include ratings for these other types of information that can be used to explain stimulus specific effects.

2. The study could be repeated with a further condition of image frequency: High frequency line drawings to isolate featural detail and low frequency blurred images to

isolate configural information. The same featural manipulations could be made and tested in both frequency conditions. If the results from the low frequency images yield lower miss rates than for the high frequency images, it would suggest that configural information is still being used even when featural detail, as in the current study, has been manipulated.

3. The study could be repeated, using a combination of compositors, given the appropriate training, to balance out compositing bias. One could hypothesise that miss rates would be higher and that perhaps the difference between features and the degree of feature replacement may be larger. However, one potential pitfall of this method could be the increase in a change in configural information through the bigger deformations of features.

4. It would be interesting to test the role of individual target features when placed into a new composite. Based on research on the whole-face context (Tanaka and Farah, 1993), it is likely that this will be a difficult task and that any memory for a specific feature will be overridden by the new whole-face context (Davies et al., 1977). This becomes relevant in practical scenarios, as some compositors may choose to use one feature from six different faces. However, this does not maximise a facial photograph database in the way that the current study does.

7.6 Conclusion

The study aimed to investigate if featural compositing, to create new faces, concealed the identity of the respective donor faces. Donor anonymity is important when creating faces for forensic applications. Two applications of created faces were the motivation for the study: creating new faces to behave as avatars in fake social media profiles, for the covert surveillance on online criminal activity and; Facial depictions, which require sampling facial features from photographs. Two experiments investigated whether any facial features were more important to be replaced than others and how many facial features need to be replaced to conceal identity. The following conclusions were made in relation to the hypotheses:

1. A facial feature saliency was found in phase 1 when individual features are replaced. The eyes demonstrated the highest saliency, followed by the hair, mouth, outline, nose and eyebrows.
2. An ordinal increase in miss rates during feature replacement was found in phase 2 for the combined data. This result was also found for the celebrity data, and mostly for lecturers. Criterion points were found for the combined data in phase 2 suggesting at least two features need to be replaced for miss rates to increase significantly (or 3 for the higher familiarity celebrity group). Further feature replacement was paired, suggesting that further replacement of one feature makes little difference, in comparison to two.
3. A main effect of feature was found for the combined data during compound feature replacement in phase 2 where eyes were significantly more salient than mouth. An interaction showed that the effect of feature was particular to replacements 2-5, where it seems to align with the criterion points of significant increases in miss rates. One interaction for the celebrity data was found at replacement 4, where higher miss rates were found for when more salient features had been replaced for one configuration, compared to when less salient features had been replaced. This was revealed as the upper and lower parts of the face (upper: including eyes and hair; lower: including mouth and outline).

Results also showed that familiarity type (celebrities and lecturers) affected feature saliency hierarchies with an internal feature preference for the higher familiarity celebrities. Overall mean miss rates were higher for the less familiar lecturers in phase 1i, where single features were replaced. However, in phase 2, celebrities yielded higher miss rates suggesting the higher familiarity level resulted in compound feature replacement being more impactful to the whole-face context. This could be attributed to an increased level of holistic processing that is normally associated with a higher level of familiarity. Identity was not concealed 100% of the time at any point in feature replacement, even when all six features were replaced, as would have been expected. Therefore, further changes need to be made, such as configural and contrast manipulations, in order to try and conceal donor identities in composites.

Contribution to theory

The results from this PhD study have supported previous findings in the literature: The feature saliency hierarchy observed in phase 1 is similar to that of previous research on face recognition, where eyes are considered the most salient (Haig, 1986, Fraser et al., 1990) (see section 2.3 Feature saliency hierarchy for a review. The different levels of familiarity, in terms of their preference for external or internal features in the literature was mostly observed between the celebrity and lecturer versions, although hair still high in saliency for the celebrity version. Normally, however, the external feature preference is for unfamiliar or newly familiar faces. This research shows that even if participants can correctly recall the identity of the lecturer, they are still displaying an external feature preference. This supports the theory, by Clutterbuck and Johnston (2007), that familiarity is on a spectrum rather than being binary.

Phase 2 yielded some surprising results for replacement 6, where some items were correctly recalled, even though all features had been replaced. This implies some residual detail was available from the face and in line with previous literature (Dakin and Watt, 2009, Schwaninger et al., 2002), participants used the other streams of information (configural and contrast) to facilitate recognition. However, this result may have, in part, been driven by the similarity of replacement features to their targets, due to the compositing technique, as mentioned already. Previous literature on the composite face effect demonstrates that the combination of two different face halves makes recognition difficult, but not impossible (Young et al., 1987). This study found that at least half of the features (n=3, celebrity data) need to be replaced for a significant increase in miss rates to occur, signifying a theoretical criterion point from old to new identity. In contrast, two changes were required in the combined data, which was experimentally more powerful, however the celebrity data yielded an approaching significant difference for subset 1 that, with more data, would potentially shift replacement 2 completely into subset 2, similar to the combined data. For celebrity data, replacement 4 showed revealed that if the upper half of the face is replaced (containing more salient eyes and hair) miss rates were significantly higher than if the lower half of the face was replaced (containing less salient outline and nose). However, further feature replacement did increase miss rates significantly but never concealed the target identity 100% of the time.

Contribution to practice

This PhD study set out to establish if compositing conceals identity of face feature/part donors so that new identities can be created for use in forensic contexts. The two forensic contexts outlined in the study were:

1. Compositing features from face databases available to investigating authorities to generate face avatars for use in online criminal investigations. How much and which parts of these faces can be used whilst still concealing the donors identity?
2. Feature sampling from face photographs for the ‘texturing’ of facial depictions would benefit from guidelines as to how much and which parts of the face can be used so that the resulting depiction does not resemble the donor.

Results have shown that the compositing technique used in the current study did not, under any of the conditions, conceal identity 100% of the time. When all six features were replaced in phase 2, some correct recognitions were given, suggesting that featural manipulations alone, are not sufficient to completely conceal identity. However, it should be noted that this result may in part be due to some residual information being available from the stimulus (e.g. configural and contrast), but also due to the potentially biased compositing technique where replacement features were perhaps not different enough to the target face, due to trying to maintain configural and contrast information.

The practical guidance for compositing for forensic scenarios, based on this study and a decision to reference the higher familiarity celebrity data-set, is to replace at least half of the face (three features) so that there is a strong chance that identity will be concealed for some of the population. Further manipulations of other streams of information, such as configural and contrast detail, will also need to be manipulated to aid in concealing identity. Given the cue of a face pool from which to extract a face memory (celebrity or lecturer) in the current study, one could hypothesise that the results found here (miss rates) could be generalised to a real-world scenario (where no cue is given) with a more generous effect. There is one caveat, however: neither types of familiarity represent the broad spectrum of familiarity in a real world scenario, even when combined. Therefore, if the results are to be generalised to other levels of familiarity, such as family members or close friends, it needs to be carried out with caution and with the suggestion that more manipulations will be needed to try and conceal identity with an even higher familiarity group than celebrities.

For the context of this study, this translates to only ever sampling no more than half the face when compositing to generate facial avatars or to sample textures from for facial depictions. And in addition, when compositing, the configuration and contrast information will most likely need to be altered.

8 References

ABUDARHAM, N. & YOVEL, G. 2016. Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision,* 16**,** 40-40.

AITPAYEV, K. & GABER, J. 2012. Creation of 3D human Avatar using Kinect. *Asian Transactions on Fundamentals of Electronics, Communications and Multimedia,* 01.

BAENNINGER, M. 1994. The development of face recognition: featural or configurational processing? *J Exp Child Psychol,* 57**,** 377-96.

BAKER, K. A., LAURENCE, S. & MONDLOCH, C. J. 2017. How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition,* 161**,** 19-30.

BARTLETT, J. C. & ABDI, H. 2006. What are the routes to face recognition? *In:* PETERSON, M. & RHODES, G. (eds.) *Perception of faces, objects, and scenes: Analytic and Holistic Processes.* Oxford: Oxford University Press.

BARTON, J. J., KEENAN, J. P. & BASS, T. 2001. Discrimination of spatial relations and features in faces: effects of inversion and viewing duration. *Br J Psychol,* 92**,** 527-49.

BARTON, J. J., RADCLIFFE, N., CHERKASOVA, M. V., EDELMAN, J. & INTRILIGATOR, J. M. 2006. Information processing during face recognition: the effects of familiarity, inversion, and morphing on scanning fixations. *Perception,* 35**,** 1089-105.

BBC. 2010. *Spanish MP's photo used in Osama Bin Laden poster* [Online]. BBC. Available: http://news.bbc.co.uk/1/hi/world/americas/8463657.stm [Accessed 2017].

BENSON, P. J. & PERRETT, D. I. 1991. Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology,* 3**,** 105-135.

BENSON, P. J. & PERRETT, D. I. 1994. Visual processing of facial distinctiveness. *Perception,* 23**,** 75-93.

BINDEMANN, M., JENKINS, R. & BURTON, A. M. 2007. A bottleneck in face identification - Repetition priming from flanker images. *Experimental Psychology,* 54**,** 192-201.

BINDEMANN, M., SCHEEPERS, C. & BURTON, A. M. 2009. Viewpoint and center of gravity affect eye movements to human faces. *J Vis,* 9**,** 7.1-16.

BLANZ, V. & VETTER, T. 1999. A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99***,** 187-194.

BONNER, L., BURTON, A. M. & BRUCE, V. 2003. Getting to know you: How we learn new faces. *Visual Cognition,* 10**,** 527-536.

BOTWIN, M. D., BUSS, D. M. & SHACKELFORD, T. K. 1997. Personality and mate preferences: five factors in mate selection and marital satisfaction. *J Pers,* 65**,** 107-36.

BOURNE, V. J., VLADEANU, M. & HOLE, G. J. 2009. Lateralised repetition priming for featurally and configurally manipulated familiar faces: Evidence for differentially lateralised processing mechanisms. *Laterality,* 14**,** 287-299.

BRIGHAM, J. C. & MALPASS, R. S. 1985. The Role of Experience and Contact in the Recognition of Faces Of Own- and Other-Race Persons. *Journal of Social Issues,* 41**,** 139-155.

BROOKS, K. R. & KEMP, R. I. 2007. Sensitivity to Feature Displacement in Familiar and Unfamiliar Faces: Beyond the Internal/External Feature Distinction. *Perception,* 36**,** 1646-1659.

BRUCE, H. L. V. 1998. Local and Relational Aspects of Face Distinctiveness. *The Quarterly Journal of Experimental Psychology A,* 51**,** 449-473.

BRUCE, V. 1982. Changing faces: visual and non-visual coding processes in face recognition. *Br J Psychol,* 73**,** 105-16.

BRUCE, V., BURTON, A. M. & DENCH, N. 1994. What's distinctive about a distinctive face? *Q J Exp Psychol A,* 47**,** 119-41.

BRUCE, V., HANNA, E. & DENCH, N. 1992. The importance of mass in line drawings of faces. *Applied Cognitive*.

BRUCE, V., HENDERSON, Z., NEWMAN, C. & BURTON, A. M. 2010. Matching identities of familiar and unfamiliar faces caught on CCTV images. *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 identities of faces. *Perception,* 23**,** 803-22.

BRUCE, V. & YOUNG, A. 1986. Understanding face recognition. *British Journal of Psychology,* 77**,** 305-327.

BRUCE, V. & YOUNG, A. W. 1998. *In the eye of the beholder: the science of face perception,* Oxford, England, Oxford University Press.

BURKE, D., TAUBERT, J. & HIGMAN, T. 2007. Are face representations viewpoint dependent? A stereo advantage for generalizing across different views of faces. *Vision research,* 47**,** 2164-9.

BURTON, A. M., BRUCE, V. & HANCOCK, P. J. B. 1999. From Pixels to People: A Model of Familiar Face Recognition. *Cognitive Science,* 23**,** 1-31.

BURTON, A. M., BRUCE, V. & JOHNSTON, R. A. 1990. Understanding face recognition with an interactive activation model. *Br J Psychol,* 81 (Pt 3)**,** 361-80.

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**,** 482-96.

BURTON, A. M., WHITE, D. & MCNEILL, A. 2010. The Glasgow Face Matching Test. *Behavior research methods,* 42**,** 286-291.

CABEZA, R. & KATO, T. 2000. Features are also important: contributions of featural and configural processing to face recognition. *Psychol Sci,* 11**,** 429-33.

CAMPBELL, R. 1999. When does the Inner-face Advantage in Familiar Face Recognition Arise and Why? *Visual Cognition,* 6**,** 197-215.

CAMPBELL, R., WALKER, J. & BARON-COHEN, S. 1995. The Development of Differential Use of Inner and Outer Face Features in Familiar Face Identification. *Journal of Experimental Child Psychology,* 59**,** 196-210.

CARBON, C.-C. & LEDER, H. 2005a. Face adaptation: Changing stable representations of familiar faces within minutes. *Advances in Experimental Psychology,* 1**,** 1-7.

CARBON, C. C. 2008. Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception,* 37**,** 801-6.

CARBON, C. C. & LEDER, H. 2005b. When feature information comes first! Early processing of inverted faces. *Perception,* 34**,** 1117-34.

CEOP 2013. Threat assessment of child sexual exploitation and abuse.

CHAI, J.-X., XIAO, J. & HODGINS, J. 2003. Vision-based control of 3D facial animation. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation.* San Diego, California: Eurographics Association.

CHELLAPPA, R., SINHA, P. & PHILLIPS, P. J. 2010. Face Recognition by Computers and Humans. *Computer,* 43**,** 46-55.

CLUTTERBUCK, R. & JOHNSTON, R. A. 2002. Exploring levels of face familiarity by using an indirect face-matching measure. *Perception,* 31**,** 985-994.

CLUTTERBUCK, R. & JOHNSTON, R. A. 2005. Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology,* 17**,** 97-116.

CLUTTERBUCK, R. & JOHNSTON, R. A. 2007. Matching as an index of face familiarity. *Visual Cognition,* 11**,** 857-869.

COLLISHAW, S. M. & HOLE, G. J. 2000. Featural and configurational processes in the recognition of faces of different familiarity. *Perception,* 29**,** 893-909.

COSATTO, E., OSTERMANN, J., GRAF, H. P. & SCHROETER, J. 2003. Lifelike talking faces for itneractive services. *Proceedings of the IEEE,* 91.

COSTEN, N. P., PARKER, D. M. & CRAW, I. 1994. Spatial content and spatial quantisation effects in face recognition. *Perception,* 23**,** 129-46.

COTE, P. 1998. FACES: The Ultimate Composite Picture (Version 3.0)[Computer software]. *Saint Hubert, Quebec, Canada: Interquest.*

DAKIN, S. C. & WATT, R. J. 2009. Biological "bar codes" in human faces. *J Vis,* 9**,** 2.1-10.

DAVIES, G., ELLIS, H. & SHEPHERD, J. 1977. Cue saliency in faces as assessed by the 'Photofit' technique. *Perception,* 6**,** 263-269.

DAVIES, G., VAN DER WILLIK, P. & MORRISON, L. J. 2000. Facial composite production: a comparison of mechanical and computer-driven systems. *J Appl Psychol,* 85**,** 119-24.

DAVIS, J. P., LANDER, K., EVANS, R. & JANSARI, A. 2016. Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers. *Applied Cognitive Psychology,* 30**,** 827-840.

DAVIS, J. P., SIMMONS, S., SULLEY, L., SOLOMON, C. J. & GIBSON, S. J. 2015. An evaluation of post-production facial composite enhancement techniques. *Journal of Forensic Practice,* 17**,** 307-318.

DE FRIAS, C. M., NILSSON, L.-G. & HERLITZ, A. 2006. Sex Differences in Cognition are Stable Over a 10-Year Period in Adulthood and Old Age. *Aging, Neuropsychology, and Cognition,* 13**,** 574-587.

DEL CARMEN, R. V. & WALKER, J. T. 2012. Chapter 14 - Lineups and other pretrial identification procedures. *Briefs of Leading Cases in Law Enforcement (Eighth Edition).* Boston: Anderson Publishing, Ltd.

DHAMECHA, T. I., SINGH, R., VATSA, M. & KUMAR, A. 2014. Recognizing Disguised Faces: Human and Machine Evaluation. *PLOS ONE,* 9**,** e99212.

EKMAN, P. 2006. *Darwin and Facial Expression: A Century of Research in Review*, Malor Books.

ELLIS, A. W., YOUNG, A. W. & FLUDE, B. M. 1990. Repetition priming and face processing: priming occurs within the system that responds to the identity of a face. *Q J Exp Psychol A,* 42**,** 495-512.

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.

EUROPOL 2017. Organised Crime Threat Assessment - Child Sexual exploitation online. *In:* CENTRE, E. C. (ed.). Europol.

FAVELLE, S. K., PALMISANO, S. & AVERY, G. 2011. Face viewpoint effects about three axes: The role of configural and featural processing. *Perception,* 40**,** 761-784.

FIELD, A. 2016. *Analysis of Covariance (ANCOVA)* [Online]. Andy Field. [Accessed 2018].

FRASER, I. H., CRAIG, G. L. & PARKER, D. M. 1990. Reaction time measures of feature saliency in schematic faces. *Perception,* 19**,** 661-673.

FREIRE, A., LEE, K. & SYMONS, L. A. 2000. The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception,* 29**,** 159-170.

FREIWALD, W. A., TSAO, D. Y. & LIVINGSTONE, M. S. 2009. A face feature space in the macaque temporal lobe. *Nature neuroscience,* 12**,** 1187-1196.

FROWD, C. 2015. Facial Composites and Techniques to Improve Image Recognizability. *In:* VALENTINE, T. & DAVIS, J. P. (eds.) *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses. Composites and CCTV.* UK: John Wiley and Sons.

FROWD, C., BRUCE, V., MCINTYRE, A. & HANCOCK, P. 2007a. The relative importance of external and internal features of facial composites. *British Journal of Psychology,* 98**,** 61-77.

FROWD, C., BRUCE, V., ROSS, D. & MCINTYRE, A. 2007b. An application of caricature: How to improve the recognition fo facial composites. *Visual Cognition,* 15**,** 954-984.

FROWD, C. D., JONES, S., FODARELLA, C., SKELTON, F. C., FIELDS, S., WILLIAMS, A., MARSH, J., THORLEY, R., NELSON, L., GREENWOOD, L., DATE, L., KEARLEY, K., MCINTYRE, A. &

HANCOCK, P. J. 2014. Configural and featural information in facial-composite images. *Science and Justice,* 54**,** 215-227.

FROWD, C. D., PITCHFORD, M., BRUCE, V., JACKSON, S., HEPTON, G., GREENALL, M., MCINTYRE, A. H. & HANCOCK, P. J. B. 2011. The Psychology of Face Construction: Giving Evolution a Helping Hand. *Applied Cognitive Psychology,* 25**,** 195-203.

FROWD, C. D., SKELTON, F., HEPTON, G., HOLDEN, L., MINAHIL, S., PITCHFORD, M., MCINTYRE, A., BROWN, C. & HANCOCK, P. J. B. 2013. Whole-face procedures for recovering facial images from memory. *Science and Justice,* 53**,** 89-97.

GAGGIOLI, A., MANTOVANI, F., CASTELNUOVO, G., WIEDERHOLD, B. & RIVA, G. 2003. Avatars in clinical psychology: a framework for the clinical use of virtual humans. *Cyberpsychol Behav,* 6**,** 117-25.

GALLAGHER, P. 2015. *Scotland yard's paedophile unit: Meeting the police men and women doing the most difficult work imaginable* [Online]. The Independent. Available: http://www.independent.co.uk/news/uk/crime/scotland-yards-paedophile-unit-meeting-the-police-men-and-women-doing-the-most-difficult-work-a6679241.html [Accessed October 2017 2017].

GALTON, F. 1883. *Inquiries into the human faculty and its development*, JM Dent and Company.

GANEL, T. & GOSHEN-GOTTSTEIN, Y. 2002. Perceptual integrality of sex and identity of faces: further evidence for the single-route hypothesis. *J Exp Psychol Hum Percept Perform,* 28**,** 854-67.

GAUTHIER, I., KLAIMAN, C. & SCHULTZ, R. T. 2009. Face composite effects reveal abnormal face processing in Autism spectrum disorders. *Vision Res,* 49**,** 470-8.

GILAD-GUTNICK, S., YOVEL, G. & SINHA, P. 2012. Recognizing degraded faces: The contribution of configural and featural cues. *Perception,* 41**,** 1497-1511.

GILAD, S., MENG, M. & SINHA, P. 2009. Role of ordinal contrast relationships in face encoding. *Proc Natl Acad Sci U S A,* 106**,** 5353-8.

GOFFAUX, V. 2012. The discriminability of local cues determines the strength of holistic face processing. *Vision Res,* 64**,** 17-22.

GOSHEN-GOTTSTEIN, Y. & GANEL, T. 2000. Repetition priming for familiar and unfamiliar faces in a sex-judgment task: evidence for a common route for the processing of sex and identity. *J Exp Psychol Learn Mem Cogn,* 26**,** 1198-214.

HAIG, N. D. 1984. The effect of feature displacement on face recognition. *Perception,* 13**,** 505-12.

HAIG, N. D. 1985. How faces differ--a new comparative technique. *Perception,* 14**,** 601-15.

HAIG, N. D. 1986. High-resolution facial feature saliency mapping. *Perception,* 15**,** 373-86.

HANCOCK, P., BRUCE, V. & BURTON, A. 2000. Recognition of unfamiliar faces. *Trends in cognitive sciences,* 4**,** 330-337.

HARRIS, A. & AGUIRRE, G. K. 2010. Neural tuning for face wholes and parts in human fusiform gyrus revealed by FMRI adaptation. *J Neurophysiol,* 104**,** 336-45.

HASEL, L. E. & WELLS, G. L. 2007. Catching the Bad Guy: Morphing Composite Faces Helps. *Law and Human Behavior,* 31**,** 193-207.

HENDERSON, J. M., WILLIAMS, C. C. & FALK, R. J. 2005. Eye movements are functional during face learning. *Mem Cognit,* 33**,** 98-106.

HENRIKSSON, L., MUR, M. & KRIEGESKORTE, N. 2015. Faciotopy—A face-feature map with face-like topology in the human occipital face area. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior,* 72**,** 156-167.

HERLITZ, A. & LOVÉN, J. 2013. Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition,* 21**,** 1306-1336.

HERLITZ, A., REUTERSKIÖLD, L., LOVÉN, J., THILERS, P. P. & REHNMAN, J. 2013. Cognitive Sex Differences Are Not Magnified as a Function of Age, Sex Hormones, or Puberty Development During Early Adolescence. *Developmental Neuropsychology,* 38**,** 167-179.

HILL, H. & BRUCE, V. 1996. Effects of lighting on the perception of facial surfaces. *Journal of experimental psychology. Human perception and performance,* 22**,** 986-1004.

HILL, H., CLAES, P., CORCORAN, M., WALTERS, M., JOHNSTON, A. & CLEMENT, J. 2011. How Different is Different? Criterion and Sensitivity in Face-Space. *Frontiers in Psychology,* 2.

HOLE, G., GEORGE, P., EAVES, K. & RAZEK, A. 2002. Effects of geometric distortions on face recognition performance. *Perception,* 31**,** 1221-1240.

HOLE, G. J. 1994. Configurational factors in the perception of unfamiliar faces. *Perception,* 23**,** 65-74.

HOLE, G. J. & BOURNE, V. 2010. *Face Processing: Psychological, Neuropsychological, and Applied perspectives,* USA, Oxford University Press.

HOLE, G. J., GEORGE, P. A. & DUNSMORE, V. 1999. Evidence for holistic processing of faces viewed as photographic negatives. *Perception,* 28**,** 341-59.

HOMMES, N. T. D. 2013a. *Sweetie* [Online]. Available: http://terredeshommesnl.org/en/sweetie [Accessed].

HOMMES, T. D. 2013b. Webcam Child Sex Tourism - Becoming Sweetie: a novel approach to stopping the global rise of Webcam Child Sex Tourism.

JARUDI, I. N. & SINHA, P. 2003. Relative Contributions of Internal and External Features to Face Recognition. Massachusetts Institute of Technology, Tech.Rep.

JIANG, F., BLANZ, V. & ROSSION, B. 2011. Holistic processing of shape cues in face identification: Evidence from face inversion, composite faces, and acquired prosopagnosia. *Visual Cognition,* 19**,** 1003-1034.

JOHNSTON, A., HILL, H. & CARMAN, N. 1992. Recognising faces: effects of lighting direction, inversion, and brightness reversal. *Perception,* 21**,** 365-75.

JOHNSTON, R. A. & EDMONDS, A. J. 2009. Familiar and unfamiliar face recognition: a review. *Memory (Hove, England),* 17**,** 577-96.

JONES, B. C., DEBRUINE, L. M., LITTLE, A. C., CONWAY, C. A., WELLING, L. L. M. & SMITH, F. 2007. Sensation seeking and men's face preferences. *Evolution and Human Behavior,* 28**,** 439-446.

KEIL, M. S. 2009. "I look in your eyes, honey": internal face features induce spatial frequency preference for human face processing. *PLoS Comput Biol,* 5**,** e1000329.

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**,** 37-52.

KLEISNER, K., PRIPLATOVA, L., FROST, P. & FLEGR, J. 2013. Trustworthy-Looking Face Meets Brown Eyes. *PLOS ONE,* 8**,** e53285.

KLONTZ, J. C. & JAIN, A. K. 2013. A Case Study of Automated Face Recognition: The Boston Marathon Bombings Suspects. *Computer,* 46**,** 91-94.

KOK, R., TAUBERT, J., VAN DER BURG, E., RHODES, G. & ALAIS, D. 2017. Face familiarity promotes stable identity recognition: exploring face perception using serial dependence. *Royal Society Open Science,* 4.

LAGUESSE, R. & ROSSION, B. 2013. Face perception is whole or none: disentangling the role of spatial contiguity and interfeature distances in the composite face illusion. *Perception,* 42**,** 1013-1026.

LANDER, K., CHRISTIE, F. & BRUCE, V. 1999. The role of movement in the recognition of famous faces. *Memory & cognition,* 27**,** 974-85.

LANGLOIS, J. H. & ROGGMAN, L. A. 1990. Attractive Faces Are Only Average. *Psychological Science,* 1**,** 115-121.

LE GRAND, R., MONDLOCH, C. J., MAURER, D. & BRENT, H. P. 2001. Neuroperception: Early visual experience and face processing. *Nature,* 410**,** 890-890.

LEDER, H. 1999. Matching person identity from facial line drawings. *Perception,* 28**,** 1171-5.

LEDER, H. & CARBON, C. C. 2005. When context hinders! Learn-test compatibility in face recognition. *Q J Exp Psychol A,* 58**,** 235-50.

LEE, K., BYATT, G. & RHODES, G. 2000. Caricature effects, distinctiveness, and identification: testing the face-space framework. *Psychological science : a journal of the American Psychological Society / APS,* 11**,** 379-85.

LEOPOLD, D. A., O'TOOLE, A. J., VETTER, T. & BLANZ, V. 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci,* 4**,** 89-94.

LEWIS, M. 2004. Face-space-R: Towards a unified account of face recognition. *Visual Cognition,* 11**,** 29-69.

LITTLE, A. C., BURT, D. M. & PERRETT, D. I. 2006. What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences,* 41**,** 1107-1118.

LITTLE, A. C., JONES, B. C. & DEBRUINE, L. M. 2011. The many faces of research on face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 366**,** 1634-1637.

LIU, J., LI, J., FENG, L., LI, L., TIAN, J. & LEE, K. 2014. Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex,* 53**,** 60-77.

LOCHTEFELD, J. G. 2002. *The Illustrated Encyclopedia of Hinduism: A-M*, Rosen.

LOGAN, A. J., GORDON, G. E. & LOFFLER, G. 2017. Contributions of individual face features to face discrimination. *Vision Research,* 137**,** 29-39.

LONGMORE, C. A., LIU, C. H. & YOUNG, A. W. 2015. The importance of internal facial features in learning new faces. *Q J Exp Psychol (Hove),* 68**,** 249-60.

LOVEN, J., HERLITZ, A. & REHNMAN, J. 2011. Women's own-gender bias in face recognition memory. *Exp Psychol,* 58**,** 333-40.

LUCAS, T. & HENNEBERG, M. 2015. Are human faces unique? A metric approach to finding single individuals without duplicates in large samples. *Forensic Sci Int,* 257**,** 514.e1-6.

LYONS, M., PLANTE, A., JEHAN, S., INOUE, S. & AKAMATSU, S. Avatar creation using automatic face recognition. ACM Multimedia 98, 1998 Bristol. 427-434.

MALCOLM, G. L., LANYON, L. J., FUGARD, A. J. & BARTON, J. J. 2008. Scan patterns during the processing of facial expression versus identity: an exploration of task-driven and stimulus-driven effects. *J Vis,* 8**,** 2.1-9.

MAN, T. W. & HILLS, P. J. 2016. Eye-tracking the own-gender bias in face recognition: Other-gender faces are viewed differently to own-gender faces. *Visual Cognition,* 24**,** 447-458.

MAROTTA, J. J., GENOVESE, C. R. & BEHRMANN, M. 2001. A functional MRI study of face recognition in patients with prosopagnosia. *Neuroreport,* 12**,** 1581-7.

MARTELLOZZO, E. 2013. *Online Child Sexual Abuse: Grooming, Policing and Child Protection in a Multi-Media World*, Taylor & Francis.

MAURER, D., GRAND, R. L. & MONDLOCH, C. J. 2002. The many faces of configural processing. *Trends in cognitive sciences,* 6**,** 255-260.

MCKONE, E. & YOVEL, G. 2009. Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychon Bull Rev,* 16**,** 778-97.

MCNEILL, D. 1999. *The Face*, Hamish Hamilton.

MEGREYA, A. M. & BURTON, A. M. 2006. Unfamiliar faces are not faces: evidence from a matching task. *Mem Cognit,* 34**,** 865-76.

MEISSNER, C. A. & BRIGHAM, J. C. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law,* 7**,** 3-35.

METZGER, M. M., KRISTOF, V. L. & YOEST, D. J. 2003. The world wide web and the laboratory: a comparison using face recognition. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society,* 6**,** 613-21.

MICHEL, C., ROSSION, B., HAN, J., CHUNG, C. S. & CALDARA, R. 2006. Holistic processing is finely tuned for faces of one's own race. *Psychol Sci,* 17**,** 608-15.

MONDLOCH, C. J., ELMS, N., MAURER, D., RHODES, G., HAYWARD, W. G., TANAKA, J. W. & ZHOU, G. 2010. Processes underlying the cross-race effect: An investigation of holistic, featural, and relational processing of own-race versus other-race faces. *Perception,* 39**,** 1065-1085.

MONDLOCH, C. J. & MAURER, D. 2008. The Effect of Face Orientation on Holistic Processing. *Perception,* 37**,** 1175-1186.

MULLINS, J. 2012. Age progression and regression. *In:* WILKINSON, C. & RYNN, C. (eds.) *Craniofacial Identification.* Cambridge: Cambridge University Press.

NASANEN, R. 1999. Spatial frequency bandwidth used in the recognition of facial images. *Vision Res,* 39**,** 3824-33.

NATU, V. & O'TOOLE, A. J. 2011. The neural processing of familiar and unfamiliar faces: a review and synopsis. *Br J Psychol,* 102**,** 726-47.

NPIA 2009. Facial identification Guidance. *In:* AGENCY, A. O. C. P. O. N. P. I. (ed.). NPIA.

O'DONNELL, C. & BRUCE, V. 2001. Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception,* 30**,** 755-764.

O'TOOLE, A. J. & PHILLIPS, P. J. 2015. Evaluating Automatic Face Recognition Systems with Human Benchmarks. *In:* VALENTINE, T. & DAVIS, J. P. (eds.) *Forensic Facial Identification.* UK: Wiley Blackwell.

OLSZEWSKA, J. I. 2016. Automated Face Recognition: Challenges and Solutions. *In:* RAMAKRISHNAN, S. (ed.) *Pattern Recognition - Analysis and Applications.* Rijeka: InTech.

PARMAR, D. N. & MEHTA, B. B. 2013. Face Recognition Methods & Applications. *International Journal of Computer Tecnology and Applications,* 4**,** 84-86.

PETERSON, M., ABBEY, C. & ECKSTEIN, M. 2007. Information distribution for face identificaiton and its relation to human strategies. *Journal of Vision,* 7**,** 121-121.

PETERSON, M., COX, I. & ECKSTEIN, M. 2008. The use of the eyes for human face recognition explained through information distribution analysis. *Journal of Vision,* 8**,** 894-894.

PEZDEK, K., O'BRIEN, M. & WASSON, C. 2012. Cross-race (but not same-race) face identification is impaired by presenting faces in a group rather than individually. *Law and Human Behavior,* 36**,** 488-495.

PHILLIPS, P. J., BEVERIDGE, J. R., DRAPER, B. A., GIVENS, G., TOOLE, A. J. O., BOLME, D. S., DUNLOP, J., LUI, Y. M., SAHIBZADA, H. & WEIMER, S. An introduction to the good, the bad, &amp; the ugly face recognition challenge problem. Face and Gesture 2011, 21-25 March 2011 2011. 346-353.

PHILLIPS, P. J. & O'TOOLE, A. J. 2014. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing,* 32**,** 74-85.

RAMON, M., BUSIGNY, T. & ROSSION, B. 2010. Impaired holistic processing of unfamiliar individual faces in acquired prosopagnosia. *Neuropsychologia,* 48**,** 933-44.

RAMON, M. & VAN BELLE, G. 2016. Real-life experience with personally familiar faces enhances discrimination based on global information. *PeerJ,* 4**,** e1465.

RHEE, C. H. & LEE, C. H. 2013. Cartoon-like Avatar generation using facial component matching. *International journal of Multimedia and ubiquitous engineering,* 8.

RHODES, G. 1988. Looking at faces: first-order and second-order features as determinants of facial appearance. *Perception,* 17**,** 43-63.

RHODES, G. 1996. *Superportraits: caricatures and recognition*, Hove: Psychology Press.

RHODES, G. 2006. The evolutionary psychology of facial beauty. *Annu Rev Psychol,* 57**,** 199-226.

RHODES, M. G. & ANASTASI, J. S. 2012. The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychol Bull,* 138**,** 146-74.

RING, T. 2016. Humans vs machines: the future of facial recognition. *Biometric Technology Today,* 2016**,** 5-8.

ROARK, D. A., ABDI, H. & O'TOOLE, A. J. 2006. When does an unfamiliar face become familiar? The effect of image type and familiarity on recognition from novel viewing conditions. *Journal of Vision,* 6**,** 12-12.

ROARK, D. A., O'TOOLE, A. J. & ABDI, H. 2003. Human Recognition of Familiar and Unfamiliar People in Naturalistic Video. *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures.* IEEE Computer Society.

ROBERT, M. B. L. 1999. A Unified Account of the Effects of Caricaturing Faces. *Visual Cognition,* 6**,** 1-42.

ROBERTS, T. & BRUCE, V. 1988. Feature saliency in judging the sex and familiarity of faces. *Perception,* 17**,** 475-481.

RUSSELL, R., SINHA, P., BIEDERMAN, I. & NEDERHOUSER, M. 2006. Is pigmentation important for face recognition? Evidence from contrast negation. *Perception,* 35**,** 749-59.

SADR, J., JARUDI, I. & SINHA, P. 2003. The role of eyebrows in face recognition. *Perception,* 32**,** 285-293.

SANCHEZ DEL RIO, J., MOCTEZUMA, D., CONDE, C., MARTIN DE DIEGO, I. & CABELLO, E. 2016. Automated border control e-gates and facial recognition systems. *Computers & Security,* 62**,** 49-72.

SANDFORD, A. & BURTON, A. M. 2014. Tolerance for distorted faces: Challenges to a configural processing account of familiar face recognition. *Cognition,* 132**,** 262-268.

SCHWANINGER, A., LOBMAIER, J. S. & COLLISHAW, S. M. 2002. Role of Featural and Configural Information in Familiar and Unfamiliar Face Recognition. *In:* BÜLTHOFF, H. H., WALLRAVEN, C., LEE, S.-W. & POGGIO, T. A. (eds.) *Biologically Motivated Computer Vision: Second International Workshop, BMCV 2002 Tübingen, Germany, November 22–24, 2002 Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

SCHWARTZ, L. & YOVEL, G. 2016. The roles of perceptual and conceptual information in face recognition. *J Exp Psychol Gen,* 145**,** 1493-1511.

SCHYNS, P. G., BONNAR, L. & GOSSELIN, F. 2002. Show me the features! Understanding recognition from the use of visual information. *Psychol Sci,* 13**,** 402-9.

SEGOVIA, K. Y., BAILENSON, J. N. & LEONETTI, C. 2012. Virtual human identification line-ups. *In:* WILKINSON, C. & RYNN, C. (eds.) *Craniofacial Identification.* Cambridge: Cambridge University Press.

SERGENT, J. 1984. An investigation into component and configural processes underlying face perception. *Br J Psychol,* 75 (Pt 2)**,** 221-42.

SERGENT, J. 1986. Microgenesis of Face Perception. *In:* ELLIS, H. D., JEEVES, M. A., NEWCOMBE, F. & YOUNG, A. (eds.) *Aspects of Face Processing.* Dordrecht: Springer Netherlands.

SHEEHAN, M. J. & NACHMAN, M. W. 2014. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature Communications,* 5**,** 4800.

SINGH, A., PATIL, D., REDDY, G. M. & OMKAR, S. N. 2017. Disguised Face Identification (DFI) with Facial KeyPoints using Spatial Fusion Convolutional Network. *CoRR,* abs/1708.09317.

SINHA, P., BALAS, B., OSTROVSKY, Y. & RUSSELL, R. 2006. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE,* 94**,** 1948-1962.

SOFER, C., DOTSCH, R., WIGBOLDUS, D. H. J. & TODOROV, A. 2014. What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science,* 26**,** 39-47.

SOLOMON, C., GIBSON, S. & MAYLIN, M. 2012. EFIT-V. *In:* WILKINSON, C. & RYNN, C. (eds.) *Craniofacial Identification.* Cambridge: Cambridge University Press.

SOUZA, F., DE LAS CASAS, D., FLORES, V., YOUN, S., CHA, M., QUERCIA, S. & ALMEIDA, V. 2015. Dawn of the Selfie Era: The Whos, Wheres, and Hows of Selfies on Instagram. *Proceedings of the 2015 ACM on Conference on Online Social Networks.* Palo Alto, California, USA: ACM.

STEEDE, L. L. & HOLE, G. J. 2006. Repetition priming and recognition of dynamic and static chimeras. *Perception,* 35**,** 1367-82.

TANAKA, J. W. & FARAH, M. J. 1993. Parts and wholes in face recognition. *The Quarterly journal of experimental psychology. A, Human experimental psychology,* 46**,** 225-45.

TANAKA, J. W. & GAUTHIER, I. 1997. Expertise in object and face recognition. *In:* GOLDSTONE, R. L., SCHYNS, P. G. & MEDIN, D. L. (eds.) *Psychology of Learning and Motivation.* San Diego: Academic Press.

TANAKA, J. W. & SENGCO, J. A. 1997. Features and their configuration in face recognition. *Memory & cognition,* 25**,** 583-92.

TANAKA, J. W. & SIMONYI, D. 2016. The "parts and wholes" of face recognition: A review of the literature. *Q J Exp Psychol (Hove)***,** 1-14.

THORNHILL, R. & GANGESTAD, S. W. 1999. Facial attractiveness. *Trends Cogn Sci,* 3**,** 452-460.

TIDDEMAN, B. 2012. Computer-generated face models. *In:* WILKINSON, C. & RYNN, C. (eds.) *Craniofacial Identification.* Cambridge: Cambridge University Press.

TONG, F. & NAKAYAMA, K. 1999. Robust representations for faces: evidence from visual search. *J Exp Psychol Hum Percept Perform,* 25**,** 1016-35.

TREMLETT, G. 2011. *Spanish MP to sue FBI for using his face in al-Qaida 'most wanted' photos* [Online]. Madrid: The Guardian. Available: https://www.theguardian.com/world/2011/oct/11/spanish-mp-fbi-al-qaida [Accessed 16/10/2017 2017].

TROJE, N. F. & BULTHOFF, H. H. 1998. How is bilateral symmetry of human faces used for recognition of novel views? *Vision Res,* 38**,** 79-89.

TURK, M. A. & PENTLAND, A. P. Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3-6 Jun 1991 1991. 586-591.

TWINSORNOT, M. *Twinsornot* [Online]. Available: twinsornot.net [Accessed 2015].

UNODC 2014. Study Facilitating the identification, description and evaluation of the effects of new information technologies on the abuse and exploitation of children. *In:* JUSTICE, C. P.

A. C. (ed.) *World crime trends and emerging issues and responses in the filed of crime prevention and criminal justice.* Vienna.

VALENTINE, T. 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A,* 43**,** 161-204.

VALENTINE, T. 2001. Face-space models of face recognition. *In:* WENGER, M. J. & TOWNSEND, J. T. (eds.) *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges.* Lawrence Erlbaum Associates Inc.

VALENTINE, T. & BRUCE, V. 1986. The effects of distinctiveness in recognising and classifying faces. *Perception,* 15**,** 525-35.

VALENTINE, T. & ENDO, M. 1992. Towards an exemplar model of face processing: the effects of race and distinctiveness. *Q J Exp Psychol A,* 44**,** 671-703.

VERES-INJAC, B. & PERSIKE, M. 2009. Recognition of briefly presented familiar and unfamiliar faces. *Psihologija,* 42**,** 47-66.

VESKER, M. & WILSON, H. R. 2012. Face Context Advantage Explained by Vernier and Separation Discrimination Acuity. *Frontiers in Psychology,* 3**,** 617.

WANG, R., LI, J., FANG, H., TIAN, M. & LIU, J. 2012. Individual differences in holistic processing predict face recognition ability. *Psychol Sci,* 23**,** 169-77.

WELLS, L. J., GILLESPIE, S. M. & ROTSHTEIN, P. 2016. Identification of Emotional Facial Expressions: Effects of Expression, Intensity, and Sex on Eye Gaze. *PLoS ONE,* 11**,** e0168307.

WHITE, M. 2004. Removing Eyebrows Impairs Recognition of Famous Faces, or Doesn't, Depending on How the Eyebrows are Removed. *Perception,* 33**,** 1215-1220.

WIESE, H., KOMES, J. & SCHWEINBERGER, S. R. 2013. Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual Cognition,* 21**,** 1337-1363.

WILKINSON, C. & EVANS, R. 2009. Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice,* 49**,** 191-196.

WILLENBOCKEL, V., SADR, J., FISET, D., O HORNE, G., GOSSELIN, F. & TANAKA, J. 2010. *Controlling low-level image properties: The SHINE toolbox*.

WILMER, J. B., GERMINE, L., CHABRIS, C. F., CHATTERJEE, G., WILLIAMS, M., LOKEN, E., NAKAYAMA, K. & DUCHAINE, B. 2010. Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences,* 107**,** 5238-5241.

WRIGHT, D. B. & SLADDEN, B. 2003. An own gender bias and the importance of hair in face recognition. *Acta Psychologica,* 114**,** 101-114.

XIE, L., JIA, J., MENG, H., DENG, Z. & WANG, L. 2015. Expressive talking avatar synthesis and animation. *Multimedia Tools and Applications,* 74**,** 9845-9848.

YIN, R. K. 1970. Face recognition by brain-injured patients: a dissociable ability? *Neuropsychologia,* 8**,** 395-402.

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**,** 737-46.

YOUNG, A. W., HELLAWELL, D. & HAY, D. C. 1987. Configurational information in face perception. *Perception,* 16**,** 747-59.

YOVEL, G. & KANWISHER, N. 2004. Face perception: domain specific, not process specific. *Neuron,* 44**,** 889-98.

ZHAO, W., CHELLAPPA, R., PHILLIPS, P. J. & ROSENFELD, A. 2003. Face recognition: A literature survey. *ACM Comput. Surv.,* 35**,** 399-458.

9 Appendices information

All appendices can be found in separate .PDF files. Each appendix document is labelled as below;

Appendix A *Methodological considerations*

Further detail and descriptions of other methodological considerations for the experimental design.

Appendix B *Unique Composites*

Description of how the unique composites were made.

Appendix C *Compositing technique*

Description of the compositing technique used to make both the unique composites and the composite stimuli.

Appendix D *MATLAB scripts*

MATLAB scripts for the assessment and matching of targets and unique composites.

Appendix E *Testing*

Screenshots of the different tasks from the experimental platforms.

Appendix F *Pilot Studies*

Full method and results of Pilot study 1 and 2.

Appendix G *Ethics*

All of the ethical approval confirmations, participant information sheets, consent forms, questionnaire and adverts. [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]

Appendix H Celebrity Targets

Images of all of the celebrity targets included in the main experiments. [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to copyright]

Appendix I Example Stimulus set

An example stimulus set from phase 1i (celebrity) that would be seen by a participant. [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to copyright]

Appendix J SPSS Output

SPSS output from the analysis for all experiments. [The appendix mentioned here cannot be made freely available via LJMU E-Theses Collection due to privacy restrictions]