Ensemble Coding of Faces

Laura Hunt

School of Psychology

University of Kent

A thesis submitted for the degree of Ph.D. in the School of Psychology at the University of Kent.

2025

Abstract

Research has consistently shown that observers are able to extract a summary of visual input. For instance, observers are able to extract an average of low-level visual features, such as the mean size and colour of a set of objects. More recently, research has investigated whether the same could occur during face perception, such as the processing of facial identity. The majority of evidence suggests that facial averages are not recognised more frequently than the individual face members of a learning set, raising doubt as to whether an internal average of these identities is actually encoded. Instead, it is possible that facial averages simply resemble the encoded set members sufficiently to be recognised as a familiar face. To investigate this possibility, the experiments in this thesis explored the ability of observers to detect the similarity of a facial average to its constituent identities. This was done by comparing the recognition of an average against its constituent exemplars under optimised conditions where set size was reduced to a logical minimum, and by using both same and different images at encoding and recognition. The exemplars and averages were also put into direct competition by asking participants to decide which face most resembled the learning set in a two-alternative forced-choice (Chapter 2). These experiments demonstrate that the average and exemplar are consistently identified at a comparable rate. The visual similarity of these stimuli was then assessed further by investigating whether participants are able to identify which faces were used to create an average (Chapter 3). Observers found this task challenging, and so a more consistent measure of similarity – facial recognition algorithms – were used to determine how much similarity an average has to its constituent identities (Chapter 4). This showed that an average created from four different faces contains minimal similarity to its constituent identities, and below a level that human observers can easily detect. These results are discussed in the context of ensemble coding and point to a cognitive mechanism that supports the formation of an internal face average.

Acknowledgements

I would like to thank my supervisor Professor Markus Bindemann for his guidance and advice throughout the PhD. I am extremely grateful for his patience and support, as well as his unique insight. I would also like to thank my friends and family for their encouragement and belief in me, especially to Jonah and Will for always being there no matter what, and to Louisa Gwynne and Oliver Herdson who either kept me sane or went crazy with me along the way. I could not have done this without all your support – you have made this a truly unforgettable experience.

Declaration

I declare that this thesis is my own work carried out under the normal terms of supervision.

Laura Hunt

Table of contents

.

CHAPTER 1:

General Introduction

1.1. Introduction

Ensemble coding refers to the encoding of an average representation of a set of stimuli and is sometimes referred to as a form of cognitive summary statistics. It is a well-established phenomenon that is consistently found for low-level properties of stimuli, such as motion direction, size, and colour (e.g., Ariely, 2001; Epstein & Emmanouil, 2017; Rajendran et al., 2021; Sweeny et al., 2013; Sweeny et al., 2015). For instance, after viewing a *set* of dots varying in size, participants are more likely to report a dot that is the *average* size of the set as having been present than a dot taken directly from the set (Ariely, 2001). This process is rapid, occurring within as little as 50 ms after viewing the set (Haberman & Whitney, 2009).

The same phenomenon also occurs for some higher-level features, such as facial expressions. For example, participants presented with sets of 4 or 16 faces varying in emotional intensity are able to determine whether a test face is more emotional than the mean of the set. This results in comparable accuracy to judging whether the test face was more emotional than a single display of emotion (Haberman & Whitney, 2007). Recent work has investigated whether this would also occur for identity. Whilst emotional expressions are relatively consistent across faces, identity is face specific. Therefore, if an identity average can be encoded, this could provide insight into how faces are learned and cognitively represented.

Current research on identity ensemble coding shows that an average identity is often categorized as an individual member of a learning set (Bagaïni & Hole, 2017; Bai et al., 2015; Ji & Hayward, 2021; Leib et al., 2014; Matthews et al., 2018; Neumann et al., 2013; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018; Sama et al., 2019). Moreover, averages are often accepted as a previously seen face at a comparable rate to individual set members (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Sama et al., 2019). However, it is unclear whether the ensemble is recognised based on its visual similarity to the learning set or because it reflects an encoded summary representation of those faces.

Research on eyewitness identification and forensic face matching shows, for example, that the identification of newly learned or unfamiliar faces – as is the case with the face stimuli employed in ensemble-coding paradigms – relies heavily on the similarity of faces (e.g., Flowe & Ebbesen, 2007; Fysh & Bindemann, 2023; Wells, 1993). Recognition of an average test face as a previously learned face may therefore reflect its visual similarity to the learned set of identities, rather than an internal cognitive representation that is an ensemble. However, as sets become larger, the identity signal for each contributing identity must be diluted. For instance, averages made from 30 images of the *same* identity are rated as having less likeness to a person than a specific exemplar image of that individual (Balas et al., 2023). It therefore follows also that an average made from *different* identities will have reduced similarity to each exemplar as additional identities are added into the ensemble.

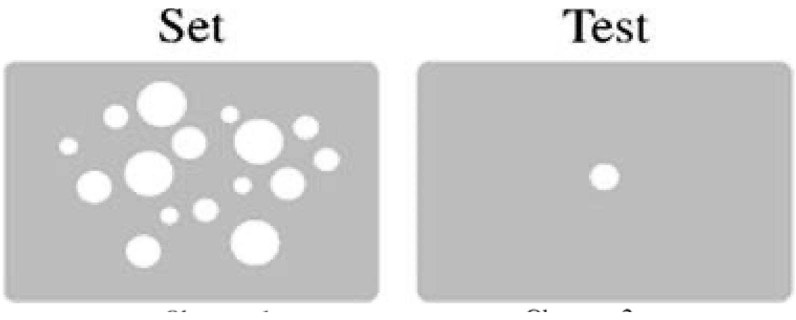
To investigate the question of whether ensemble coding reflects the formation of internal average-based representations or only the processing of visual similarity between faces, this thesis will explore how much similarity an average made from multiple identities contains to each of its constituent identities, by investigating the ability of observers to detect this resemblance. The first aim is to examine recognition of an average against its constituent exemplars under optimised conditions, with reduced set sizes and using both identical and different images of the same identities at learning and test. The second aim concerns the ability to detect the resemblance between individual set members and the average face. The final aim is to explore how facial similarity relates to ensemble coding by comparing similarity scores generated by a facial comparison algorithm for the exemplar and average faces against the set members. This thesis begins by discussing ensemble coding of low-level visual features before outlining ensemble coding of facial information. Current accepted evidence for the ensemble coding of identity will be reviewed. This chapter will end with a general description of the approach this research will take.

1.2. Ensemble Coding of Low-Level Visual Features

The phenomenon of ensemble coding was first discovered within motion research, whereby a set of dots moving in different directions was still perceived as having coherent motion based on the average (Williams & Sekuler, 1984). It has since become a well-established phenomenon that is consistently found for low-level properties of stimuli, such as size and colour (e.g., Ariely, 2001; Epstein & Emmanouil, 2017; Rajendran et al., 2021; Sweeny et al., 2013; Sweeny et al., 2015). For instance, a seminal study demonstrated that average size could be extracted from a set of dots (Ariely, 2001). This study presented participants with a set of 4, 8, 12, or 16 dots varying in size (see Figure 1.1). There were four sizes of dots presented in each display, but the frequency of each size varied, with no more than four of one size per trial. These were presented for 500 ms. After this, participants were presented with a test stimulus comprising a single dot. This dot was either a member of the set, a new dot, or the mean size of the set. The task was to report whether the test was a member of the set or not. Results showed that the closer the test stimulus was to the mean of the set, the more likely participants were to report the dot to have been in the set. These findings have since been replicated repeatedly, demonstrating that observers can extract an ensemble of size (e.g., Chong & Treisman, 2003; Chong & Treisman, 2005).

Figure 1.1

An example of the stimuli used by Ariely (2001). The top left panel shows the learning set and the top right panel the test stimulus. Observers have to decide whether the test stimulus appeared in the learning set.



Ensemble effects have since been found for other low-level visual properties. This includes colour (Rajendran et al., 2021), length (Kacin et al., 2021), orientation (Kacin et al., 2021), and shape (Elias & Sweeny, 2020). Ensembles of separate low-level features can be formed in parallel. For instance, an ensemble average of colour can be encoded at the same time as an ensemble of orientation (Hansmann-Roth et al., 2019). Similarly, multiple ensembles can be encoded for the same feature. For example, two colour ensembles can be encoded simultaneously (e.g., Chong & Treisman, 2005; Utochkin, 2015). Ensemble coding therefore appears to be a highly efficient mechanism to extract visual information.

Ensemble coding is also a *rapid* process, occurring within as little as 50 ms after viewing a set (Haberman & Whitney, 2009). This suggests that ensemble coding is a heuristic used by the cognitive visual system to quickly encode large amounts of visual information. For instance, visual scenes contain considerable amounts of information, yet can be processed rapidly (e.g., Rousselet et al., 2005; Wiesmann & Võ, 2022). This is driven by the 'gist' of the scene, whereby a visual summary is extracted based on low-level properties, such as colour (Castelhano & Henderson, 2008; Furtak et al., 2022; Oliva & Schyns, 2000). Although this can facilitate recognition of objects that are consistent with the scene (e.g., Castelhano & Henderson, 2008; Furtak et al., 2022), intricate details of the scene are often not encoded. For instance, in change detection paradigms, minor changes to details within a visual scene often go unnoticed (Mäntylä & Sundström, 2004), suggesting that only a global summary of the scene is encoded. Participants who demonstrate more global ensemble coding benefit more on object naming tasks when an object is consistent with a briefly presented scene (Brady et al., 2017), suggesting ensemble coding drives scene 'gist'. This implies that ensemble coding enables the rapid extraction of the global features of visual scenes, but not the detailed processing of the individual elements within a scene.

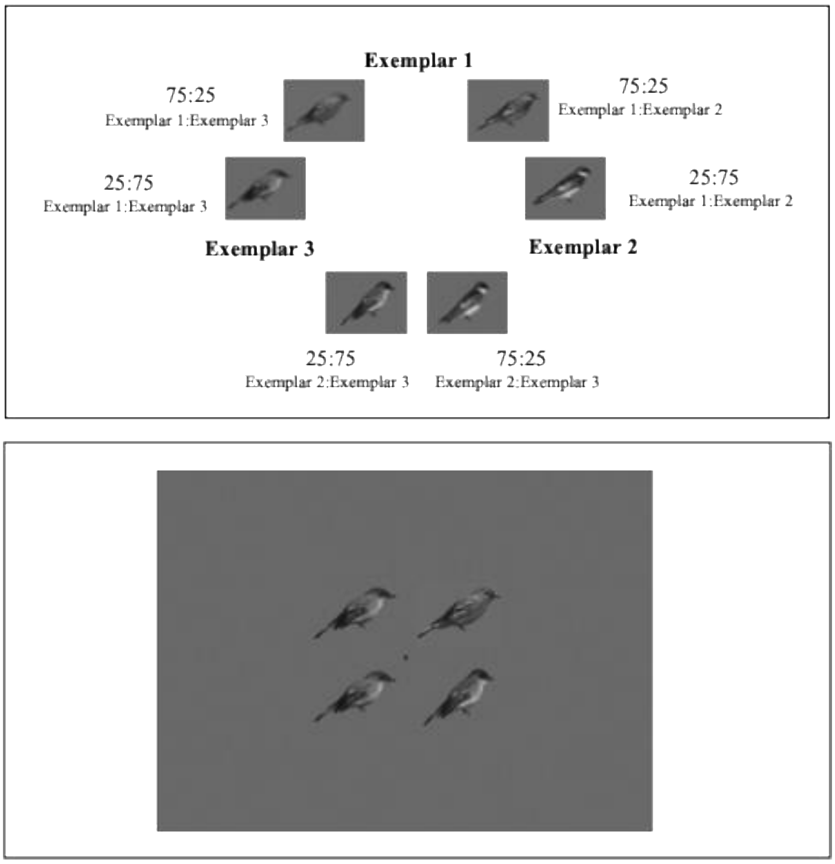
The extraction of low-level visual ensembles can be understood through well-known mechanisms of the human visual system. For instance, the colour of an object can change depending on the luminance. Despite this, the visual system is able to adjust for this variation so that the object is perceived as a consistent colour (Hurlbert, 2007). It is therefore logical that the visual system can encode the average colour to enable the perception of colour constancy. Similarly, the visual cortex contains direction-selective cells which respond when a stimulus moves in a particular direction. These project onto neurons within the middle temporal (MT) visual area. Neurons here are suppressed if there is competing movement. Therefore, the direction which has the strongest signal – that is, the *average* direction – will be perceived (see Andersen, 1997).

1.3. Ensemble Coding of High-Level Visual Features

High-level visual information can also be encoded into an ensemble. For instance, participants are able to identify the mean *object* from a set of the same category (Chang & Gauthier, 2022). In this study, participants were presented with a set of four objects from the same category, such as different birds. All objects used were morphed from three original exemplars (see Figure 2). For example, one object was a 25:75 morph of two exemplars and another was a 75:25 morph of the same exemplars. In each display, two of the set members were identical, whilst the other two were the adjacent morphs on the continuum (Figure 1.2). This meant that the repeated set members became the mean of the set. Participants were then presented with six objects and asked to identify which object was the mean of the set. Accuracy on this test was approximately 65%, suggesting participants can extract an average of specific objects.

Figure 1.2

Example stimuli used in Chang and Gauthier (2022). The top panel shows an example of a continuum of the morphed stimuli. The bottom panel shows the encoding set given to observers with four morphs. Of these four, two are the same, whilst the other two are the morphs to the left and right of the continuum.



There is consistent evidence that observers can also extract an average of facial expressions (e.g., Elias et al., 2017; Griffiths et al., 2018; Haberman & Whitney, 2009; Haberman & Whitney, 2010; Haberman & Whitney, 2011; Han et al., 2021; Karaminis et al., 2017; Li et al., 2016; Leib et al., 2012; Sun & Chong, 2020). In these paradigms, participants are typically presented with a set of images of a single face. Each of these images is morphed along a continuum from two extreme emotions (i.e., happy and sad), resulting in different intensities of emotion. Each face in the set contains a different emotional intensity. Participants are then presented with a single test face and asked whether this face is happier or sadder than the mean of the set (e.g., Haberman & Whitney, 2007) or to report if the test face was a set member (e.g., Haberman & Whitney, 2009). With the latter, participants presented with sets of 4, 8, 12, or 16 faces are more likely to select a face that represents the average emotional intensity of the set (> 70%) than either a face taken directly from a set or a face with a different emotional intensity to the set (Haberman & Whitney, 2009).

This finding has been replicated across a variety of conditions, such as with clinical samples. For instance, individuals with prosopagnosia, a condition characterised by an inability to recognise faces, show ensemble coding of emotion (Leib et al., 2012). Individuals with autism also demonstrate ensemble coding for emotion (Karaminis et al., 2017; Rhodes et al., 2015), despite difficulties in processing facial emotion (e.g., Kuusikko et al., 2009; Rump et al., 2009). Emotional ensemble coding has also been found across larger set sizes, with effects found with sets of 24 (Wolfe et al., 2015; Yang et al., 2013), short exposure times of 50 ms (Li et al., 2016), and dynamic expressions (Elias et al., 2017).

If ensemble coding occurs for dynamic facial information such as emotional expression, then this raises the question of whether similar effects are observed for more rigid facial information, such as identity. In this case, viewers presented with a set of different identities should report that an average image of these, in which the faces of several people are combined into a single representation, had been seen in the set previously. However, identity is a multidimensional concept (e.g., Valentine et al., 2016) and as more faces are added into an ensemble average, this will result in reduced similarity of the average to each contributing identity. In this review, current evidence for the ensemble coding of identity will be discussed and evaluated to understand how much similarity an average has to its constituent identities.

1.4.Potential Mechanisms of Identity Ensemble Coding

Faces show variability in appearance between different identities but also within an identity (i.e., hairstyle, make-up, viewpoint). These two sources of variability can interact, in that two images of the same person can appear very different, while two images of different people can sometimes also look remarkably alike. Despite this, familiar faces are recognised with ease, even under impoverished conditions, such as with highly blurred or pixelated images (e.g., Bruce et al., 2001; Lander & Bruce, 2001). This raises the question of how unfamiliar faces become familiar and their identification becomes robust to variability in appearance.

1.4.1. Cognitive Recoding

One possible answer to this question lies in how the cognitive system organises visual input. Evidence suggests that familiar faces may be internally represented as an average to deal with variability in facial appearance. Faces can show huge amounts of variability based on both superficial changes, such as the lighting and viewpoint from which a face is seen, as well as longer-term changes such as ageing (e.g., Ritchie & Burton, 2017; Sexton et al., 2023). For instance, just a 5-degree change in viewpoint reduces the ability to detect which of two faces were a previously seen target, with a 20-degree change resulting in 1.69 times poorer performance (Swystun & Logan, 2019). Moreover, the quality of an image can also impact the ability to extract identity-specific information on a face. For example, by reducing the bit rate of CCTV style videos from 92 to 52 kbps, hit rate falls by 18%

(Keval & Sasse, 2008). Additionally, changes to image resolution through blurring reduces the ability to accurately recognise a face (Gilad-Gutnick et al., 2012). While familiar faces are easily recognised despite these factors (e.g., Bruce et al., 2001), unfamiliar face recognition can be impeded substantially (Jenkins et al., 2011).

Unfamiliar face recognition also becomes increasingly difficult when the two instances of the face are spread over a larger period of time. For example, in border security settings, passport officers have to compare a live person to a photograph that could have been taken up to 10 years ago. This task is challenging because people show large amounts of within-person variability in facial appearance over time. For instance, changes in hair, skin tone, and weight occur over time, which can alter the appearance of a person. Hairstyle changes, for example, result in lower recognition rates even when the face itself remains identical (Bartel et al., 2018). Appearances can also vary dramatically even over short time frames. For example, the addition of glasses can have a strong influence on a person's appearance (Kramer & Ritchie, 2016). However, unfamiliar face identification remains challenging even under optimised conditions. For example, even two pictures taken on the same day just a few minutes apart, but with two different cameras, are not easily matched as the same identity (see, e.g., Burton et al., 2010; Megreya & Burton, 2006b).

Some individuals are also better able to recognise unfamiliar faces than others. These individuals are known as super-recognisers and score approximately 13% higher than typical observers (Robertson et al., 2016). It has been proposed that super-recognisers facial recognition abilities stem from a greater ability to encode within-person variability (e.g., White et al., 2022). Although super-recognisers also show greater ability to distinguish *between* identities than typical observers, this effect is weaker than their superior ability in encoding within-person variability

(White, et al., 2022). Learning the extent of this within-identity variability is therefore essential for accurate face recognition. However, within-identity variability is idiosyncratic (Burton, 2013; Burton et al., 2016; Jenkins et al., 2011) in that each identity demonstrates different degrees of this variability. Many encounters with a face are therefore required to receive sufficient exposure to its variability. This raises the issue of how this variability is cognitively represented to enable familiar faces (and unfamiliar faces for super-recognisers) to be so easily recognised.

One possibility is that each encounter of a face is stored as a separate view-dependent image. Over time, a collection of these images would build up which provided more templates to enable recognition of that face in the future. This view was grounded in research that demonstrates that exposing participants to multiple images of a single identity benefits recognition of this identity from a novel image compared to a single image or an average of all images (Andrews et al., 2015; Ritchie et al., 2017). For instance, providing participants with 20 images of an identity at learning facilitates later recognition of a novel image, particularly when the learning images are high in variability (Ritchie et al., 2017). This can also occur if the images are learned incidentally in a prior face sorting task (Andrews et al., 2015).

However, other evidence has led to the idea that this variability in faces is captured from an internal representation in the form of an average (Burton et al., 2005; Kramer et al., 2015). For example, presenting participants with an average of a famous celebrity, which is created by morphing a set of different images of that person, has been shown to result in more accurate recognition performance than using a single image of that identity (Burton et al., 2005). Similarly, averages of an unfamiliar identity can also be recognised if viewers have been exposed to multiple images of that identity (Kramer et al., 2015). One explanation for this effect is that the coding of a face into an average enables the cognitive visual system to filter out 'noise' – that is, the variation in appearance that exists across different images of the same person. This allows more identity-relevant information that is consistent across images to be kept stable. Such average encoding of facial identity resonates with theorising from the ensemble coding domain, by indicating that the cognitive system may also recode an average of different facial identities in a similar fashion.

1.4.2. Capacity Limits

Related to the notion that facial identities might be coded as averages is the apparent decline in recognition ability when there is more than one unfamiliar target (Megreya & Burton, 2006a; Nortje et al., 2017). For example, in a laboratory lineup task, in which observers have to identify a target from a subsequent lineup of ten faces, recognition accuracy for a single target has been shown to stand at 60%. However, this falls to 34% when an additional target is introduced (Megreya & Burton, 2006a). Under more naturalistic conditions, accuracy falls from 40% to 30% when identifying a person asking for directions either alone or with another person respectively and continues to decline with additional targets (Clifford & Hollin, 1981). This disadvantage can be found when the additional identities are task-irrelevant (Jenkins et al., 2003) or when part of a separate task (Palermo & Rhodes, 2002). This demonstrates that the processes that drive identity encoding must have limits on the capacity of information they can hold in working memory at once.

This capacity limit seems to be face-specific, as facial recognition under equivalent conditions is unaffected when the distractor is a non-face stimulus (Jenkins et al., 2003). In this study, participants were asked to judge whether a name belonged to a pop star or a politician. This was presented alongside a distractor face, which was either the face that matched the name or a face belonging to the opposite semantic category. A congruency effect was found, whereby reaction times were longer when the distractor was a face from the opposite semantic category. However, when an additional unknown face was presented alongside the name and distractor face, this congruency effect was reduced, suggesting the additional face was impairing processing. However, the congruency effect remained unaffected if the additional stimulus was a non-face object (Jenkins et al., 2003). This demonstrates that faces have their own independent processing capacity which is unaffected by non-face stimuli. Nevertheless, the amount of facial information that can be processed at one time is limited. Interference effects from distracting stimuli are eliminated when participants are required to learn more than one target identity (Thoma & Lavie, 2013). This points to a strict capacity limit, whereby only a single face can be *identified* at a time (Bindemann et al., 2005; Thoma & Lavie, 2013). However, participants are able to report if a display face contains all faces or a mix of face and non-face objects (Qarooni et al., 2022), suggesting participants are still able to *detect* multiple faces. This capacity limit therefore does not affect face detection, but likely occurs during the encoding of the identities (Bindemann, Sandford, et al., 2012). Taken together, these findings suggest that the cognitive system responsible for facial encoding can only process a single identity at a time.

The encoding of identity must therefore occur rapidly if humans are able to accurately encode and recognise the abundance of faces encountered daily. However, evidence from ERP studies suggests that it takes 5-10 minutes of exposure for a face to start generating ERP effects seen for familiar faces, such as the N250 (Popova & Wiese, 2023). Full familiarity with a face can take up to 14 months to develop

(Popova & Wiese, 2022). It is therefore unsurprising that humans have difficulty in recognising unfamiliar faces after brief exposure. How multiple unfamiliar identities are perceived simultaneously and encoded therefore remains uncertain. One possible explanation is that cognitive heuristics are drawn upon to process and encode large amounts of information at once, namely ensemble coding.

1.4.3. Context vs. Co-Occurrence

What information is collated into a single ensemble representation may depend on three possible mechanisms. The first is context. It is clear that multiple exposures to an identity are required to form a stable representation (e.g., Jenkins et al., 2011; Menon et al., 2015). However, two different unfamiliar identities are often mistaken as the same person. For instance, in matching tasks, whereby observers compare two images to determine if they belong to the same identity (a match) or two different identities (a mismatch), observers incorrectly accept mismatch trials as a match between 12-34% of the time (Burton et al., 2010; Fysh & Bindemann, 2018; White et al., 2014). Alongside this, recognising a matching identity is challenging, even when multiple images of that identity have been encoded, if the encoding images are labelled under different names (Menon et al., 2018).

Context therefore is essential in building this understanding of variability. Prosopagnosic patients, who have impairments in recognising familiar faces, often use contextual cues, such as clothing, location, and speech to identify the individual (e.g., Hoover et al., 2010; Portch et al., 2023). These cues can provide the cognitive system with evidence that the face matches a previously encoded identity, thus enabling the system to update the cognitive representation. It is therefore also possible that multiple identities might be *averaged* together based on related context cues. Indeed, ensemble coding becomes more precise when contextual cues, such as background, remain stable across trials (Jia et al., 2023).

The second potential mechanism is co-occurrence, where faces encountered at the same time or place might be grouped together. For instance, co-occurrence of two faces at learning facilitates associative priming, whereby priming participants with one of the faces increases recognition of the other (Vladeanu et al., 2006). In some ensemble coding paradigms, participants are presented with two sets of groups simultaneously, both consisting of members that vary in emotional intensity. Participants are then asked to select which group had the higher mean emotion. Participants can perform accurately on these tasks (Haberman & Whitney, 2011; Im et al., 2017; Sun & Chong, 2020). This shows that participants can form an ensemble based off stimuli that co-occur in a group, rather than based off all items on screen.

The final mechanism is similarity. Similarity is complex as this is not an objective measure. If similarity was objective, similarity would be based solely off physical resemblance between items. However, in an object recognition task where observers had to decide which of two objects matched a target, ambiguous shapes that fell under the same category of label (e.g., star) produced slower reaction times than shapes that fell under separate categories, despite matching similarity distances (Zettersten & Lupyan, 2020). This demonstrates that similarity also draws on more subjective measures. This is dependent on an observer's perceptual experiences and beliefs. For instance, learning items in categories can facilitate how similar items within a category are, and how dissimilar items between categories are (e.g., Foroni & Rothbart, 2011). These effects can also be clearly seen in the face literature. For example, recognition of a face is enhanced if participants believe the face to be a member of an in-group (i.e., same university) than an out-group (Bernstein et al., 2007). Similarly, two images of the same identity can be learned as the same or two separate identities depending on labelling (Menon et al., 2018).

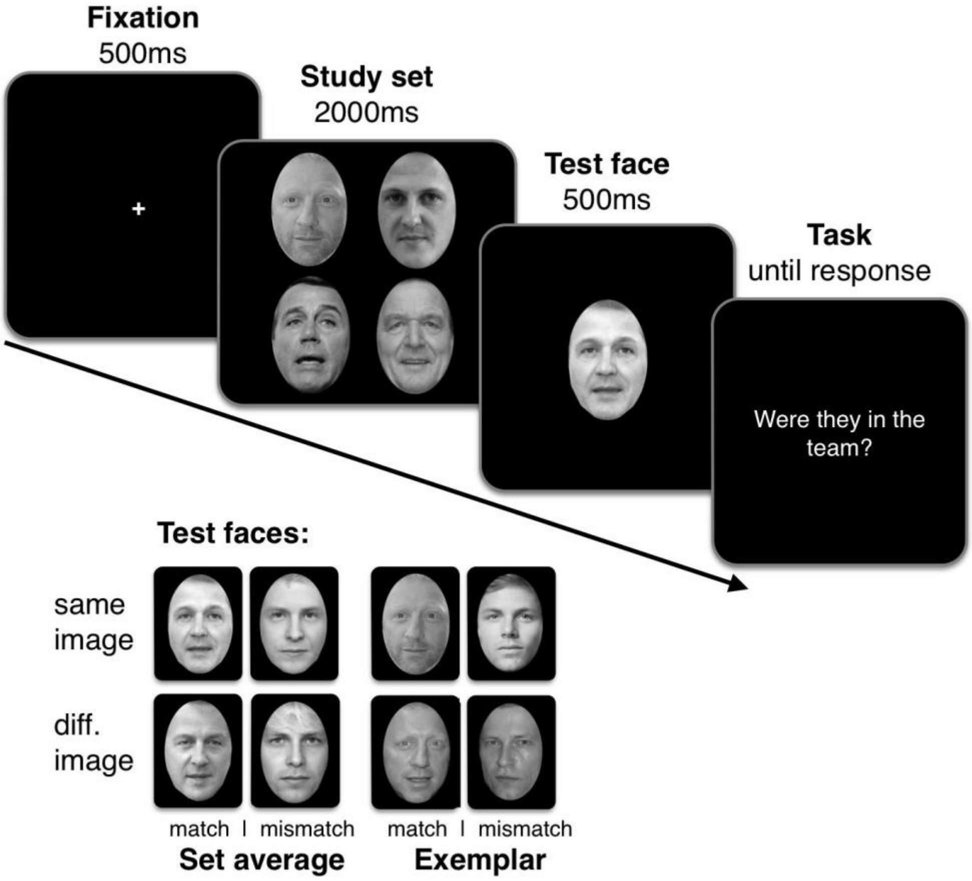
Indeed, groups of similar faces are more likely to be encoded into an ensemble than dissimilar ones (Neumann et al., 2015). However, this only applies for more objective physical similarity, as perceived subjective similarity for other-race groups does not increase ensemble effects (Peng et al., 2021). This implies that ensemble coding is an objective phenomenon and draws upon different mechanisms to face perception, which are altered to fit with an individual's beliefs and experiences.

1.5. Ensemble Coding of Identity

There are relatively few studies so far that have investigated the ensemble coding of face identity (see Table 1.1). These studies often employ a matching paradigm (Figure 1.3) in which participants are shown an encoding set of targets (the study set), ranging from two to 18 identities, for around 1000-2000 ms. Participants are then presented with a probe face (the test face). This is either an individual taken from the set (a matching exemplar), an average of all the identities in the set (a matching average), or a new identity (a mismatching exemplar), or an average created from previously unseen identities (a mismatching average). Using this type of paradigm, some studies have observed the *more frequent* selection of matching average faces than matching exemplars (e.g., Matthews et al., 2018; Peng, Kuang, et al., 2019; Rhodes et al., 2015). Additionally, non-matching averages are not selected any more frequently than non-matching exemplars (Peng, Kuang, et al., 2019; Rhodes et al., 2015).

Figure 1.3

An example of the matching paradigm as used in Rhodes et al. (2018).



However, this pattern only seems to occur under specific conditions. For example, the more frequent selection of average faces over exemplars has been found in children aged 6-18 years. For adults, however, averages were less likely to be selected than exemplars (Matthews et al., 2018). This average advantage for children has also been found in other studies, however, only for children aged 6-8-years-old (Robson et al., 2018). One of the potential reasons behind this is that younger children perceive things differently. For instance, younger children have been shown to have higher levels of holistic processing, whereby faces are processed as a whole rather than as individual parts. This has been found across a range of stimuli (e.g., faces, watches), whereas in adults, holistic processing mainly occurs for faces but no other objects (Meinhardt-Injac et al., 2017). Ensemble representations of size also tend to occur more frequently in children than adults, with adults showing more accurate exemplar discrimination (Sweeny et al., 2015). This may therefore explain why children have a stronger tendency to select the average identity over the exemplar than adults. More frequent identification of the average is also found when identification is tested across different images of the same people at learning and test in children aged 8-16 (Rhodes et al., 2015). These images were collected from an image search and varied randomly along several dimensions such as lighting and facial expression (Rhodes et al., 2015). This provides evidence that ensemble coding does occur for identity and is not a by-product of other image properties that have been found to elicit ensemble coding, such as lighting or facial emotion. Other studies have also found additional conditions that seem to elicit a preferential selection of averages in adults, for instance, when a set size of four or less is used (Peng, Kuang, et al., 2019). This casts doubt on the strength of ensemble effects, as with less identities used to create the average, the higher each identity's contribution to the average is. It is possible that this is then selected because it resembles each exemplar sufficiently to facilitate recognition.

Other studies using this paradigm have found accuracy rates for the average to be *comparable* to that of the exemplar, providing some further evidence for an ensemble representation of identity (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022). Across these studies, recognition rates for the exemplar tend to fall at approximately 60%. For the average, this is 57%, showing highly similar recognition of both the exemplar and average faces (Bagaïni & Hole, 2017; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2013; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). However, in most studies the average is selected less frequently than the exemplars, casting doubt on how much similarity to each identity is retained in an average (Ji & Hayward, 2021; Neumann et al., 2013; Neumann et al., 2018; Rhodes et al., 2018).

In addition, other studies have used an adjustment paradigm, in which a group of targets are presented in a similar fashion to the matching paradigm. However, participants are then required to adjust the subsequent probe face using a computer mouse to scroll through a continuum of faces until it matches either the average of the set or a cued individual exemplar. In these studies, the accuracy rates of the exemplar are also sometimes found to be similar to the average (Sama et al., 2019) or higher than the average (Bai et al., 2015; Leib et al., 2014). Taken together, the studies detailed here indicate that most of the time, averages are recognised less frequently than exemplars (9 studies), averages and exemplars are recognised similarly some of the time (5 studies), and rarely are averages recognised more frequently than exemplars (3 studies). These studies are summarised in Table 1.1. The variation in findings raises the question of how ensemble coding should be defined, and which conditions tend to give rise to this effect. This review aims to discuss this question in relation to what has currently been accepted as evidence for the ensemble coding of identi

Table 1.1

Current Research on the Ensemble Coding of Identity. Columns show the differences in the paradigms, including the number of encoding identities within a set and whether these were displayed simultaneously or sequentially. Also provided is whether the studies used the same or different images at encoding and recognition, the durations of the encoding set and probe face display, and how the selection of the average compared to the exemplar.

Authors (Date)	No. of Encoding Identities	Image Type	Type of Test	Display	Encoding Duration	Participant Age	Probe Face Duration	Selection of average compared to exemplar
Bagaïni & Hole (2017)	4	Same	Match/Mismatch	Sequential	500 ms	Adult	500 ms	=
Bai et al. (2015)	18	Same	Adjust to Mean	Simultaneous	1000 ms	Adult	N/A	<
Ji & Hayward (2021)	4	Same	Match/Mismatch	Simultaneous	2000 ms	Adult	Until Response	<
Leib et al. (2014)	2-18	Same	Adjust to Mean	Sequential	50-850 ms	Adult	Until Response	<

Matthews et al. (2018)	4	Same	Match/Mismatch	Simultaneous /Sequential	1500-2500 ms/375-500 ms	Adults & children aged 6-10/6-11	500-1000 ms	>
Neumann et al. (2013)	4	Same/Different	Match/Mismatch	Simultaneous	1500 ms	Adults	500 ms	<
Neumann et al. (2018)	2-8	Same	Match/Mismatch	Simultaneous	50-6400 ms	Adults	500 ms	<
Peng, Kuang, et al. (2019)	4/9	Same	Match/Mismatch	Simultaneous	2000 ms	Adults	Until Response	4 targets: > 9 targets: <
Peng, Zhang, et al. (2019)	4	Same	Match/Mismatch	Simultaneous	2000 ms	Adults	Until Response	<
Peng et al. (2021)	4	Same	Match/Mismatch	Simultaneous	2000 ms	Adults	Until Response	=
Peng et al. (2022)	4	Same	Match/Mismatch	Simultaneous	2000 ms	Adults	Until Response	=
Rhodes et al. (2015)	4	Same/Different	Match/Mismatch	Simultaneous	2000 ms	Children aged 8-16 and aged	500 ms	Same images: =

						children 9-14 with ASD		Different images: >
Rhodes et al. (2018)	4	Same/Different	Match/Mismatch	Simultaneous	2000 ms	Children aged 6-15 and adolescents aged 17-18	500 ms	<
Robson et al. (2018)	4	Same/Different	Match/Mismatch	Simultaneous	2000 ms	Adults	500 ms	<
Sama et al. (2019)	6	Same	Adjust to Mean	Simultaneous	400 ms	Adults	Until Response	=

Note. Evidence for more frequent recognition of the ensemble (>), equal recognition (=), or less frequent recognition of the ensemble (<) compared to the exemplar

1.6. More Frequent Selection of the Average

In measuring ensemble coding, a key question concerns how ensemble coding should be defined and operationalised. One argument is that a face that has not been seen previously in a set should not outperform a face from the set unless it matches the internal cognitive representations by which the set faces have been stored. Therefore, confidence that ensemble coding has been observed should be highest when averages are selected more frequently than exemplars.

1.6.1. Equal Encoding of all Faces

This view is grounded in the expectation that participants encode *all* faces in the set during the formation of an ensemble representation. However, the literature on face perception raises the question of whether this is likely to happen in practice. Face processing, for instance, is subject to capacity limits, whereby only one face identity can be processed at a time (Bindemann et al., 2005). Moreover, whilst emotion can be encoded rapidly and automatically (e.g., Codispoti et al., 2009; Jiang et al., 2009), identity may require longer exposure times. For instance, during the comparison of pairs of faces, 500 ms seems to be required to fixate on each face once, but performance is best when further fixations are possible (Özbek & Bindemann, 2011). Considering that studies of ensemble coding of facial identity typically present sets of four faces for 2000 ms (e.g., Ji & Hayward, 2021; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018), which may be sufficient for the encoding of emotion, the question arises of whether this leaves sufficient time to encode each facial identity accurately.

There is some evidence that ensemble coding is independent of attention (e.g., Alvarez & Oliva, 2008; Alvarez & Oliva, 2009; Epstein & Emmanouil, 2017; Liu & Ji, 2024; Peng, Kuang, et al., 2019). If this is the case, then it is possible that all faces can be incorporated into an ensemble quickly, without the more time-consuming encoding of all faces within a set. However, evidence is currently mixed as to whether ensemble coding depends on attentional mechanisms or not. Whilst there is evidence that attention is not required for the ensemble coding of low-level visual information, such as spatial layout (e.g., Alvarez & Oliva, 2009), there is little evidence that the same applies for faces. In fact, most studies show that attention is required for the ensemble coding of faces, particularly facial identity (Chen & Zhou, 2018; Ying, 2022). In these studies, participants' attention is directed to attend to a subset of target faces among the learning set. They are then presented with a test face comprising of one of the target exemplars, an average of the targets, an average of the whole set, or a new average. When asked to respond if the test face was one of the targets, these studies typically find more frequent endorsement of the target average than the whole set average, but comparable recognition of the target average and target exemplar (Chen & Zhou, 2018). This indicates that attention to faces is required for ensemble coding.

A related question concerns whether all faces are encoded *equally*. If time for encoding is limited, then faces that are fixated first may be encoded to great depth, whereas other faces may be encoded in less detail or not at all. This seems plausible not only given the short display time but also considering viewing biases in visual perception. For example, individuals in Western cultures often show a bias towards stimuli presented in the left visual field (Guo et al., 2009). Attention is allocated preferentially to some facial stimuli on the basis of other factors, such as attractiveness or group membership. Attractive faces, for example, are viewed for longer than unattractive ones (Leder et al., 2010), as are faces that belong to an observer's own ethnic group (Lovén et al., 2012). Indeed, participants are more likely to form an ensemble of specific targets within a set than one of the whole set, such as targets that have the same colour or pattern (Khvostov et al., 2024), suggesting attentional limits may result in encoding a subset of faces.

It is therefore possible that the faces in a learning set of an ensemble experiment are not encoded equally, but that some faces are more likely to be encoded than others, and that there is insufficient time to encode some faces at all. Such factors could then bias results to produce a pattern that might appear to be an ensemble coding effect. Consider a scenario where averages and exemplars are equally likely to be selected by participants (e.g., Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019). If there is insufficient time to encode all exemplars from a learning set due to limited display times – say, on average, only three out of four faces, then this places an upper bound of 75% accuracy that can be achieved in the exemplar condition (i.e., the exemplar that was not encoded would not be recognised). However, the average of these faces might still be recognised more frequently, for the most part, if it consists of faces that *were* encoded (3/4 of the learning set). In this case, the average condition could produce higher accuracy (i.e., 100%) than the exemplar condition (at 75%) can possibly reach. There is evidence that capacity limits, own-face, and own-gender biases are involved in the ensemble coding of faces (e.g., Ji et al., 2018; Peng et al., 2021; Thornton et al., 2019). These considerations indicate that further controls are needed to assert the presence of ensemble coding effects. One method for doing so could be to eye-track observers during encoding, to provide a measure of the extent to which all faces are being processed. However, while eye-tracking has been employed extensively in the face domain (e.g., Guo & Shaw, 2015; Hills, 2018; Man & Hills, 2016; Stacey et al., 2005; Wu et al., 2012), it has not been combined with the ensemble coding of faces to date.

A different method by which studies have attempted to control for the subsampling of identities during encoding is through the use of sequential presentations, so that every face in the learning set is viewed individually. Some studies have used sequential presentations in place of simultaneous presentations and found accuracy rates for the average to be equal to that of the exemplar, in line with previous findings (e.g., Bagaïni & Hole, 2017). Other studies have compared simultaneous and sequential encoding directly, and found no difference in ensemble coding (e.g., Matthews et al., 2018). One way in which studies have directly measured the effect of simultaneous to sequential presentation on subsampling is by varying the number of targets shown in the learning set but keeping the number of identities used to create the average the same across trials. If participants only remembered a subset of faces, then error should remain stable regardless of how many faces were in the learning set. On the other hand, if an ensemble representation is formed internally, then error should reduce as the number of faces in the learning set increases, as the internal representation of these faces would become more similar to the average image. Error was found to reduce as set size increased, suggesting that participants were not only learning a subsample of the set, but incorporating all identities into an ensemble (Leib et al., 2014). Therefore, whilst ensembles can be formed based on shared attributes between set members, viewers are also able to incorporate all set members into an ensemble of these cannot be readily categorised into distinct groups.

1.6.2. Image vs. Identity-Coding

Previous literature on face processing suggests that identity can be processed very rapidly. For instance, research on repetition priming, in which responses to a stimulus change as a result of prior exposure to that stimulus, has shown that this effect can occur across different images of the same identity (Bindemann et al., 2007; Schweinberger et al., 2002). However, image characteristics can also be rapidly encoded. For example, identity matching from two identical images of a single identity is more accurate than matching from two different images of the same identity (e.g., Jenkins & Burton, 2011). Recognition across a *change in image* is important for demonstrating that an *identity* rather than just a specific image has been processed. For this reason, participants are typically asked to identify a target across a change in image in eyewitness paradigms. When these tasks are performed with identities that were initially unfamiliar to observers, performance is prone to error (~70%; Megreya & Burton, 2006a).

This is an important issue for ensemble coding, as most studies have used the same images of the identities for both the learning set and the test face (e.g., Bagaïni & Hole, 2017; Bai et al., 2015; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019). This raises the possibility that participants are encoding an ensemble of the image characteristics, rather than the identities in the set. There have been only *four* studies that have used different images (Neumann et al., 2013; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). In two of these studies, using the same image, as opposed to different images, at learning and test resulted in the more frequent selection of averages (Rhodes et al., 2015; Robson et al., 2018). For instance, a mean proportion of .64

present responses was found for the average when a different image was used compared to .49 for the exemplar, in contrast to the .70 and .73 respectively when the same images were used (Rhodes et al., 2015). Moreover, selection of the exemplar declines more than the selection of the average when a different image is used (see Figure 3; Robson et al., 2018). This suggests that exemplar recognition may rely more on image characteristics, whereas the average is more resistant to this noise. However, the average is a face that has been *created*. Therefore, the average image will always be different to that of the individual faces used in a learning set, even if created from the same images. It is therefore not surprising that the selection frequency of the average is less affected by a change in image than the exemplar.

1.7. Any Selection of the Average

An alternative characterisation of ensemble coding is that *any* recognition of the average, even if this occurs less frequently than recognition of exemplars, is sufficient evidence for ensemble coding of identity. The argument for this view is that the average image has never been previously seen. In this case, there should not be any recognition of this face and therefore any recognition of averages arises due to the internal representation of an ensemble.

As previously mentioned, only three of the 15 studies showed more frequent selection of the average (Matthews et al., 2018; Peng, Kuang, et al., 2019; Rhodes et al., 2015). Of the remaining 12 studies, four have found averages to be selected as frequently as exemplars (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019) and eight have found the averages to be selected less frequently than exemplars (Bai et al., 2015; Ji & Hayward, 2021; Leib et al., 2014; Neumann et al., 2013; Neumann et al., 2014; Peng, Zhang, et al., 2019; Rhodes et al., 2018;

Robson et al., 2018). Given that the average will have *some* resemblance to its constituent exemplars, it would be surprising if the average was never selected, especially when considering identification must tolerate a degree of dissimilarity across different images of the same face (e.g., Corpuz & Oriet, 2022; Jenkins et al., 2011). Nevertheless, these studies reason that there is sufficient evidence for the ensemble coding of identity, under the interpretation that *any* selection of the average is evidence for recognition.

1.8. Simultaneous Encoding of the Exemplar and Average

A potential problem with the notion that any selection of the average is evidence of its recognition arises from a consideration of what the internal cognitive representation of faces might be. Understanding what form this internal representation takes is a challenging task. One way to determine what this representation is, is to determine which stimulus most resembles this. If the internal representation is an ensemble, then participants should recognise an average more, and vice versa for the exemplar, as this will maximise overlap between the encoded faces and the probe stimulus. However, neither seems to be the case, with mixed evidence as to what face, if any, is the stronger representation (Bagaïni & Hole, 2017; Matthews et al., 2018; Neumann et al., 2013). Alternatively, *both* the exemplar and the average may be encoded simultaneously. Research has found recognition of the average face to increase with longer viewing times and smaller set sizes, but the same increase was found for exemplar recognition rates (Neumann et al., 2018). This suggests that both exemplars and their average may be encoded simultaneously.

However, despite such evidence, some studies also point to an independence in these processes. For example, whilst attentional capacity places limits on the encoding of exemplars, this does not affect the average (Li et al., 2021). This is likely because averages are formed rapidly, whereas exemplars require more time to be processed. For instance, evidence suggests that ensemble representations have an advantage under shorter exposure times, whereas exemplar representations are stronger with longer exposure times (Liu et al., 2023). The N2pc has been linked to attention, particularly spatial attention, whereby attention is directed to a specific location (Galfano et al., 2011; Kiss et al., 2008; Robitaille & Jolicoeur, 2006). This N2pc component is also elicited when individual faces capture attention (Bola et al., 2021; Wieser et al., 2018; Zhou et al., 2015). This effect seems to be linked selectively to the encoding of individual faces, but not the average (Liu et al., 2023), which is likely encoded independent of spatial attention (Kacin et al., 2022). This is surprising, as it suggests the average is encoded without attention to each individual face. This points to ensemble representations acting as a 'gist', where a global summary of the identities is rapidly extracted and combined into a single ensemble representation.

Additionally, holistic processing may underlie ensemble coding more so than the coding of individual faces (e.g., Norman & Tokarev, 2014). Although face processing is thought to be a holistic process (e.g., Richler et al., 2011; Van Belle et al., 2010), for unfamiliar faces featural processing becomes more important (e.g., Lobmaier & Mast, 2007; Megreya & Burton, 2006b). However, the ensemble coding of facial information seems to rely fully on holistic processing. For instance, cultures who place a high value on interdependence, where individuals view themselves as a collective rather than as independent autonomous beings, demonstrate more holistic processing (Nisbett & Miyamoto, 2005). These individuals also demonstrate stronger ensemble coding than participants from cultures which focus on independence (Im et al., 2017).

Holistic and featural processing demonstrate considerable similarity to ensemble and exemplar coding. For instance, face processing may use both holistic and featural processing, but with holistic processing preceding featural processing (Richler et al., 2009). Evidence for this comes from research which demonstrates that holistic processing is used more frequently under short exposure times, but featural processing increases as exposure time increases (Hole, 1994). Interestingly, increased exposure time does not reduce holistic processing (Hole, 1994). This is remarkably similar to the effects seen for the exemplar and the average, whereby ensemble coding seems to precede exemplar coding, and could explain the temporal precedence of the ensemble representation over the exemplar (Liu et al., 2023). Moreover, whilst exemplar coding increases with exposure time, ensemble coding remains stable (Liu et al., 2023).

The question remains as to whether the encoding of individual exemplars is a requirement for ensemble coding. There is evidence that the differences in the mean emotion between two groups can be detected when the individual changes cannot be detected (Haberman & Whitney, 2011). Moreover, objects that are individually unrecognisable still contribute to an ensemble representation (Fischer & Whitney, 2011). Individuals with prosopagnosia, a condition that impairs the ability to recognise even familiar faces, demonstrate some level of ensemble coding, where averages are identified at a similar frequency to the exemplar, although accuracy is reduced relative to controls (Robson et al., 2018). This implies that the ability to recognise an average does not draw on the same mechanisms as the ability to recognise exemplars.

1.9. Similarity of an Average to its Exemplars

The potential simultaneous encoding of exemplars and averages poses a challenge in determining if an ensemble is actually being encoded internally. It is possible that the external average that is typically employed in ensemble paradigms as a probe stimulus, is recognised because it retains enough resemblance to the encoded exemplars to pass a decision threshold. This could also give the average face an advantage over the exemplar in situations where only a subset of exemplars are encoded. For instance, if the exemplar at test was an identity *not* encoded, then this face would not be selected. However, the average face would still contain facial information from the encoded exemplars, and if this similarity reaches the decision threshold, this face would still be selected without the encoding of an *internal* average. A fundamental question therefore concerns how much similarity an average has to its exemplars and whether individuals can perceive this similarity.

1.9.1. Face Matching

It is already clear that observers have difficulty seeing similarity in faces that are unfamiliar to them. For instance, the ability to identify an unfamiliar face drops from 90% with the same image at learning and test to 60% when a different image is used (Bruce, 1982). Whilst part of this difficulty can be attributed to general memory requirements, there is also a perceptual element to the issue. Evidence for this comes from face matching tasks, whereby observers compare two faces to determine if these belong to the same individual (a match) or two different individuals (a mismatch). These stimuli are often optimised to provide the best chance of an accurate response. For instance, face stimuli are often acquired from a frontal view and high-quality images are taken under suitable lighting conditions. Even under these optimised conditions, such as with photos taken on the same day just minutes apart, but with a different camera, identification errors of 10% are still observed (Burton et al, 2010). This effect is exacerbated further when images are taken under more ambient conditions. For example, with face photographs from student identity cards, which are free to vary in dimensions such as facial expression, lighting and image quality, accuracy rates drop to 66% (Kent Face Matching Test (KFMT); Fysh & Bindemann, 2018). This demonstrates the difficulties that observers can have in seeing the resemblance between faces, as they often struggle to simply match two images of the same identity.

1.9.2. Facial Morphing

However, there is also evidence that observers can see the resemblance between an average and its exemplars. This comes from research on facial morphs, in which one image changes gradually into another in a series of steps along a continuum. Facial morphs have been used to create fraudulent identity documents to pass at border security, where an image of two identities averaged together is used as a passport image, to enable two people to use the same passport. Studies investigating these attacks have found that facial morphs are accepted as an ambient image of one of the exemplar identities between 48-68% of the time (Kramer et al., 2019; Robertson et al., 2017). This demonstrates that an average of two identities, where the average is created from 50% of each identity, not only contains sufficient resemblance for observers to detect, but to pass observers' internal identification thresholds whereby this averaged image is frequently accepted as an ambient image of a particular identity.

However, as more identities are included within an average, the similarity to each exemplar is inevitably reduced. For instance, studies that have manipulated the contribution of each identity to a two-face average have shown that as the contribution of one exemplar decreases, the less likely it is to be accepted as a match to that exemplar (Rotshtein et al., 2005). For instance, if a morph contains only 20% of one identity, then this will typically go undetected, with the average being accepted as an ambient image of the identity comprising 80% of the average (Jenkins & Burton, 2011). Extrapolating these results to ensemble coding, this implies that averages made from learning sets of four or more faces, where each identity contribute 25% or less to the average, should be unlikely to be recognised based off resemblance to a specific exemplar.

1.10. Structure of this Thesis

The aim of this thesis is to investigate how much similarity to each exemplar is retained within an average. This will provide direct insight into whether the internal representation of a set of faces is likely to be an ensemble or whether an external average face can be recognised based on resemblance to the encoded exemplars. Chapter 2 first examines whether observers will recognise an average of two targets – the minimum number of targets needed for an ensemble effect to occur with the highest chance of resemblance. To do this, an old/new paradigm is used, whereby observers are shown an encoding set and then this is followed by a test face. Participants are asked to determine if this test face was from the encoding set or not. This is investigated under conditions where the test face is made up of the same images used at learning (Experiment 1) and different images (Experiment 2) to investigate image-based and identity processing. As no studies have yet put the exemplar and average into direct competition whereby observers are required to select which of these faces most resembles one of the previously encoded identities. Experiment 3 therefore investigates which face, the exemplar or average, participants will choose to determine which face has the strongest resemblance to the internal representation (Experiment 3).

The possibility that recognition of the average face is the result of its resemblance to an internal exemplar representation can only hold if participants can *see* the resemblance between an average and its constituent identities. Chapter 3 therefore further explores the internal representation by investigating whether participants are able to detect the resemblance between an average face and the identities that constitute the average. This is achieved by providing participants with a four-face average and four images of different identities. Participants are then asked to select any of the identities they believe were used to create the average. This is implemented as a matching task (Experiments 4 and 5) and a memory task (Experiments 6 and 7). Potential effects of memory are further explored by manipulating working memory load (Experiment 7). Two-face averages and four exemplar images are included as a comparison (Experiments 5-7) to explore whether the ability to detect resemblance changes as a function of the number of identities incorporated into an average or whether the task is simply harder with four faces to compare.

As human perception of similarity can be inconsistent, Chapter 4 explores the role of resemblance in more detail by obtaining accurate similarity ratings from a facial recognition algorithm. First, accuracy on the algorithm is established using two well-established tests of face identity matching– the Glasgow Face Matching Test (Burton et al., 2010) and the Kent Face Matching Test (Fysh & Bindemann, 2018). Similarity scores are then compared for the exemplar and average faces against the encoding images. Finally, these similarity scores are correlated against the behavioural data from Chapters 2 and 3 to determine if similarity can explain participants' recognition of the average and their ability to detect resemblance.

CHAPTER 2:

Ensemble Coding of Identity Under Optimal Conditions

Introduction

Ensemble coding refers to the integration of a set of similar stimuli into a shared cognitive representation that captures their global properties. This effect has been demonstrated with various stimulus categories, such as visual shapes that vary in orientation or size (Ariely, 2001; Sweeny et al., 2015). In these experiments, participants are typically presented with an encoding set of stimuli that vary along these dimensions. They are then shown a probe stimulus at a recognition test phase and asked to report whether this appeared in the encoding set. This stimulus would either be an item taken from the set (the exemplar), the average of the set, or a new item. In these studies, participants typically report recognition probe items, such as a circle of a specific diameter, as being from an encoding set if this falls within the range of sizes of the encoding set (e.g., a display of circles of different sizes) (Ariely, 2001; Khayat & Hochstein, 2018; Oh et al., 2019). This suggests that observers are drawing some kind of internal summary representations of the group, rather than a precise replication of the visual information.

These effects are also observed with more complex social stimuli such as faces. For example, when participants are presented with a set of faces varying in emotional intensity, a face that captures the average emotional intensity of the encoding set is often selected at recognition, even if this face has not been previously seen (Haberman & Whitney, 2009; Li et al., 2016; Goldenberg et al., 2020). These effects have been replicated under a range of conditions, such as with dynamic emotional and spatially distributed faces (Elias et al., 2017; Han et al., 2021) and appear to be remarkably robust. This raises the question of whether other facial information is also subject to ensemble coding.

An interesting candidate for such encoding is facial identity. Faces exhibit within-person variability in appearance. This reflects moment-to-moment changes, such as non-rigid facial movements from emotional expression and speech (Burton et al., 2016; Christie & Bruce, 1998), situational variables such as lighting direction and viewpoint (Hancock et al., 2000), as well as changes that occur over time due to ageing and so forth (Burton, 2013). Some theories of face recognition propose that the cognitive system deals with this variability by encoding facial identities into averages – statistical summaries of the different encounters with a person's face that extract consistencies in appearance across different instances, and in which information that does not consistently code identity is discarded (Burton et al., 2005; Jenkins & Burton, 2011). There has been evidence to support this theorising from simulations and behavioural data (Burton et al., 2005; Jenkins & Burton, 2008; Koca & Oriet, 2023; Ritchie et al., 2018). However, evidence is mixed as to whether ensemble coding is the mechanism by which such cognitive representations might be formed.

Despite a number of differences between studies, ranging from different display times during encoding of the learning set (e.g., 500 ms, 1000 ms, 6400 ms) (Bagaïni & Hole, 2017; Bai et al., 2015; Ji & Hayward, 2021; Neumann et al., 2018), sets that vary in size (Bai et al., 2015; Ji & Hayward, 2021; Neumann et al., 2018), and different paradigms (e.g., matching and mean selection tasks) (Leib et al., 2014; Rhodes et al., 2018; Robson et al., 2018; Sama et al., 2019), the average is consistently recognised more often than a new non-matching identity. However, unlike low-level ensemble coding with simple non-face objects, there is mixed evidence as to the strength of the ensemble representation over the exemplar faces from which it is formed.

Some studies show that average faces are selected more often at recognition than individual exemplars from an encoding set (e.g., Matthews et al., 2018; Rhodes et al., 2015; Peng, Kuang, et al., 2019). This resonates with theorising on face recognition, which shows that an identity average can be more than the sum of its parts, by outperforming exemplars in simulations and face-name verification tasks (e.g., Burton et al., 2005). In studies of ensemble coding, such superior performance is observed when different images of the identities are used at encoding and recognition (Rhodes et al., 2015). Such cross-image recognition indicates that this process is independent of the specific image of a person and operates at the level of identity (see Bruce, 1994; Brunas et al., 1990; Ellis et al., 1990).

However, most research has accepted that *any* recognition of the average is evidence that ensemble coding is taking place, regardless of how it compares relative to the recognition rates for the exemplar face (Ji & Hayward, 2021; Neumann et al., 2013; Neumann et al., 2018; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2018; Robson et al., 2018;). This interpretation is based on the argument that the average is a face that has never been encountered before. Therefore, any 'recognition' of this face would only occur if it matched an internal representation that was formed from multiple exemplars during face encoding.

A key issue to consider here may be that the average contains some identity information from all exemplars seen at encoding. This face should therefore contain some resemblance to each exemplar. It is possible that this resemblance is sufficient to pass a decision threshold, where the average is then accepted as a match to the encoded exemplars, rather than an encoded *ensemble*. This makes it difficult to conclude that an internal ensemble representation exists based on *any* recognition of the average. If faces are stored as an average, then recognition should be higher for such an image than exemplars, as it is a direct match with the internal cognitive representation. Whilst only a few studies find evidence of more frequent recognition of the average (e.g., Matthews et al., 2018; Peng, Kuang, et al., 2019; Rhodes et al., 2015), the differences in the manipulation of stimuli and implementation of ensemble coding paradigms hampers comparison of the relative strength of exemplars and averages.

Another way to investigate whether an internal ensemble is formed is to compare the recognition rates for the average when participants encode a single target to when they have encoded multiple targets. In ensemble paradigms, participants are shown sets of faces ranging from two to 18 identities (e.g., Bagaïni & Hole, 2017; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2013; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). As yet, no studies have investigated recognition of an average created from two faces when participants only encode one of the exemplars. Under these conditions, the average face would comprise a single encoded target and a new face. If recognition of the average is based on resemblance to the encoded exemplars, then recognition of the average should be comparable when encoding either one or two of the targets, as the resemblance of the average to a single exemplar would be the same. However, if observers are forming an internal ensemble, then the average should only be selected at the recognition phase if participants have encoded multiple targets. However, no studies have thus far investigated whether the average is recognised when participants only encode a single identity.

To examine this possibility, the current experiments investigated this ensemble coding in its most basic form. In contrast to previous studies, the encoding set was reduced to a logical minimum of two identities. Each encoding display was then followed by a recognition phase consisting of a single probe face. This face either comprised of an exemplar (one of the two encoded faces), an average (a morph of the two encoded faces), or a new identity. Participants were then asked to report whether the probe face is a match to these target(s) or not. The question of main interest here was whether an average can outperform exemplar recognition. If the average face is recognised more frequently than an exemplar, this would point to an internal average. As it is also possible that average and exemplar conditions might produce equivalent performance, a further comparison condition was provided, in which an average was created from a single encoded exemplar and a new face. As this stimulus contains information not encoded, but the same amount of resemblance to a single encoded exemplar, recognition of this face would point to the average being selected due to its resemblance to a *single* encoded exemplar face rather than the formation of an internal ensemble of *two* faces.

Experiment 1

The purpose of this experiment was to determine whether face averages would be recognised more frequently than exemplar faces when multiple targets are encoded in a paradigm that reduces ensemble coding to its most basic form. On each trial of this experiment, participants either encoded one face in the single-face condition, or a pair of faces in the two-face condition. They were then presented with a probe face and asked to determine if this was a match to the encoding set. In the single-face condition, the probe face was either an exemplar from the encoding set or an average of this face and a previously unseen (new) face (non-matching average). This provides an important control condition: when only a single face is encoded, if the average is selected because it *resembles* the encoded identity, the average should be identified as frequently in this condition as it is in two-face conditions as the resemblance to a specific exemplar remains the same. In contrast, in the two-face condition, the probe face was either one of the encoded exemplars or an average of both encoded exemplars (matching average). The question of main interest here was whether recognition for an average of two learned faces would be more frequent than for its constituent exemplars and if this occurs specifically when multiple targets are encoded, indicating an internal ensemble representation.

Methods

Participants

Based on previous sample sizes used in research on multiple-face identifications (Bindemann, Sandford, et al., 2012; Megreya & Burton, 2006b). thirty-four participants (5 male, 22 female), with a mean age of 24.7 years (*SD* = 8.7), took part in the experiment in exchange for course credit at the University of Kent or via Prolific for a small fee.

Stimuli

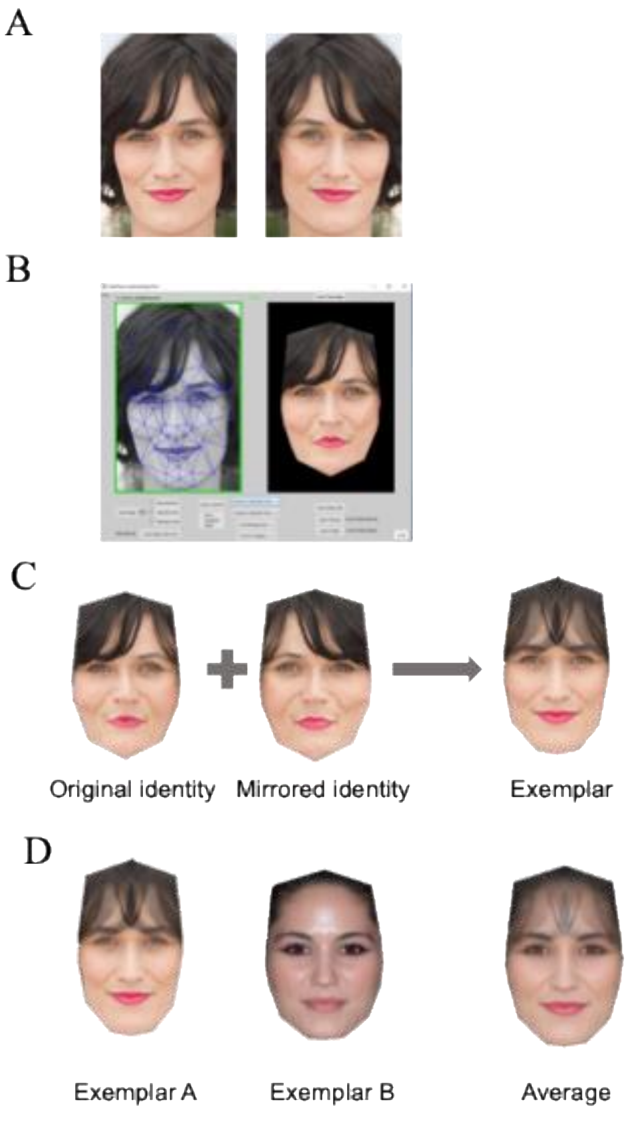
This experiment employed a 2 (Set Size: One vs. Two) x 3 (Face Type: Exemplar vs. Average vs. New) repeated-measures design. The face stimuli for this experiment were sourced from an internet search of French and German celebrities, so that multiple images of each identity could be obtained, but these identities were expected to be unfamiliar to UK participants. Two-hundred-and-forty (120 male, 120 female) identities were chosen. For each identity, a Google Image search was run, and three images were selected that showed the face in a frontal view. For this experiment, only one of the three images was used, while the other two images were collected for follow-up experiments. These identities were then paired up randomly with the restriction that both faces were of the same sex and from the same country, resulting in 120 pairings. On trials on which only a single face was shown during the encoding phase, only one of the two identities within each pairing was utilised as a target, whereas the other face served as a new identity on other trials.

To ensure that exemplars and averages could not be distinguished along superficial dimensions (e.g., enhanced symmetry) or artefacts of the averaging process (e.g., the blurring of wrinkles and hair), exemplars were also presented as averages at recognition. These averaged exemplars were created using the same images as encoding by duplicating each face and flipping this along the horizontal axis (Figure 2.1). The original and mirror-reversed exemplar were then combined as an average using InterFace software (see Kramer, Jenkins, & Burton, 2017). For this, the key features of each face (eyes, nose, mouth, outline) were landmarked with a polygon mesh which enables the extraction of shape and texture information from each image. Shape information refers to the positions of key features of the face, such as the eyes, nose, and mouth. Texture information of the face includes information about lighting and reflectance and once extracted, is stored on a standard face template. Once these data have been extracted for both the original and mirror-reversed exemplar, the shape and texture information from the images being morphed were then averaged together. This results in a single face that contained the average shape and texture of the two images – the exemplar. Two-hundred-and-forty exemplar averages were created from the 120 pairs of identities. One-hundred-and-twenty of these were used as exemplars, with the other 120 used as new identities.

Finally, for the two-face condition, ‘ensemble’ averages were created by averaging together the two averaged exemplar identities in each pair using the same method, to create a total of 360 stimuli (120 exemplars, 120 averages, and 120 new faces). This gave rise to 120 different trials, of which half were one-face trials and the other half-two face trials, and comprised of 40 exemplar tests, 40 average trials, and 40 new trials. Each encoding face was shown at a size of 160 (w) x 217 (h) pixels, with a resolution of 54 ppi. All probe faces were shown at approximately 118 (w) x 190 (h) pixels with a resolution of 54 ppi.

Figure 2.1

An illustration of how stimuli were created starting with the mirroring (A) and landmarking of the images (B) and then the averaging of the images to create the exemplar (C). Two exemplars were then combined in the same manner to create the cross-identity average (D).



Procedure

The study was programmed on PsychoPy (Version 2.8; Peirce et al., 2019) and data was collected online using Prolific.com. On each trial, a fixation cross appeared for 1000ms, followed by the encoding phase, which comprised either of a single face for 2250 ms or two faces for 4500 ms based on previous literature on multiple target identifications (Bindemann, Sandford, et al., 2012). The different exposure times were employed to equate the available inspection time *per* face. After the presentation of the encoding phase, a second fixation cross was shown for 1000 ms, followed by the recognition phase for 200 ms. Participants were told to press 'D' if the probe face was a match to the single target or one of the two targets or to press 'L' if the probe was a new face that they had not seen before. There was no time limit for participants to complete the task, but participants were told to respond as accurately as possible. Encoding set size and probe face type were presented in a randomly intermixed order, but the appearance of the different identities was counterbalanced across experimental conditions over the course of the experiment. Participants were given a short break after 60 trials.

On completion of this task, participants completed a familiarity check to exclude any identities that they might know. This involved the individual viewing of all 240 identities in the experiment. Alongside each face, four names were presented. Participants were asked to select the correct name of the celebrity by entering the corresponding number or to press 'L' if the face was familiar, but they could not recall the name. If participants did not recognise the identity, they were asked to press 'S'. The experiment took approximately 45 minutes to complete.

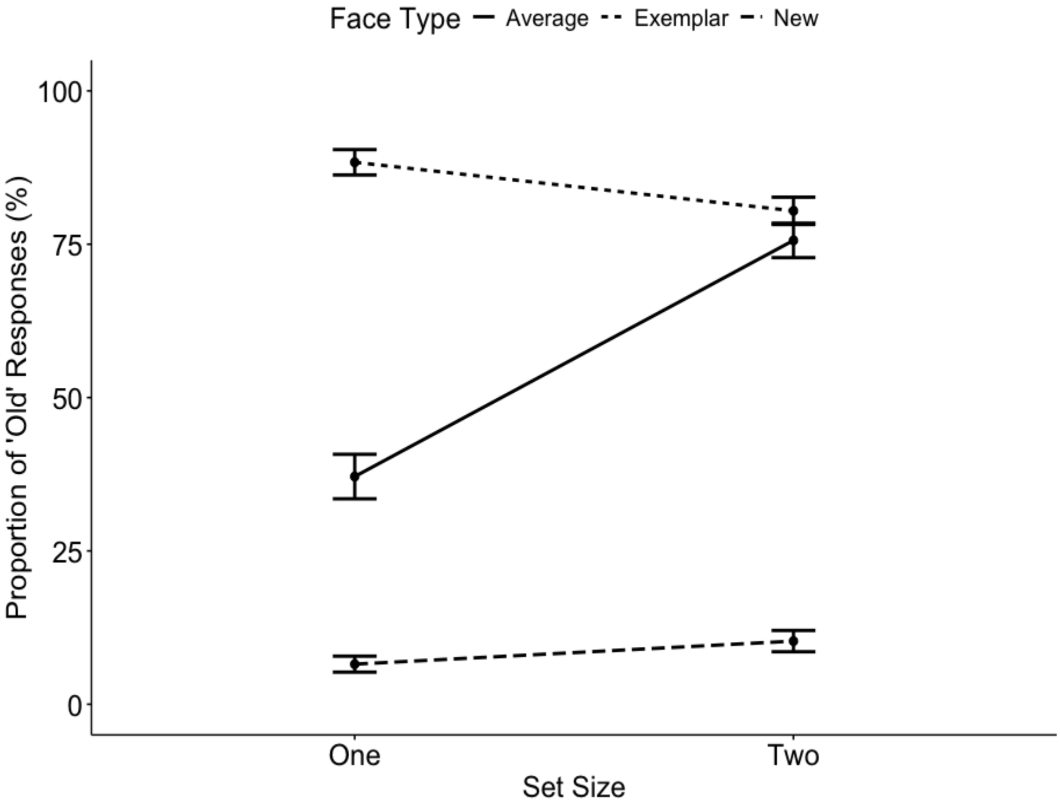
Results

Any trials which included a face that a participant reported as familiar during the familiarity check was removed from the analysis. This resulted in the removal of 482 trials, equating to 11.8% of all trials. The percentage of 'old' of responses for all remaining trials was then analysed. The proportion of old responses for new faces was low, at 6.5% (*SD* = 7.6) and 10.3% (*SD* = 10.0) for the one- and two-face conditions, indicating that these identities were rarely mistaken as faces from the encoding set. The data of primary interest comprised of the percentage of 'old' responses for exemplar and average faces and can be seen in Figure 2.2. A 2 (Set Size: One vs. Two) x 3 (Face Type: Exemplar vs. Average vs. New) repeated-measures ANOVA of this data revealed a main effect of set size, $F(1, 33) = 58.52$, $p < .001$, $\eta_p^2 = .64$, with more 'old' responses on two-face ($M = 78.03$, $SD = 14.86$), compared to one-face trials ($M = 62.64$, $SD = 30.95$), and a main effect of face type, $F(2, 66) = 487.48$, $p < .001$, $\eta_p^2 = 94$, due to more 'old' responses on exemplar ($M = 84.40$, $SD = 13.04$), $p < .001$, and average trials ($M = 56.37$, $SD = 26.99$), $p < .001$, than on new trials ($M = 8.42$, $SD = 9.08$). In addition, more 'old' responses were also made to exemplar faces than average faces, $p < .001$. These effects were qualified by an interaction between set size and face type, $F(2, 66) = 90.04$, $p < .001$, $\eta_p^2 = .73$.

Tukey's HSD test was conducted to analyse this interaction. This revealed more 'old' responses for both exemplar ($M = 88.36$, $SD = 12.05$), $p < .001$, and average faces ($M = 37.12$, $SD = 21.15$), $p < .001$, compared to new faces ($M = 6.54$, $SD = 7.59$) in the one-face conditions. In addition, the proportions of 'old' responses were lower for averages than exemplars in the one-face conditions, $p < .001$, as the average comprises of an encoded and an unseen identity. In contrast, exemplars ($M = 80.44$, $SD = 12.96$) and averages ($M = 75.62$, $SD = 16.38$) were classified as 'old' with comparable frequency on two-face trials, $p = .261$, with both exemplars, $p < .001$, and averages, $p < .001$, being classified as 'old' more frequently than new faces ($M = 10.30$, $SD = 10.11$) on two-face trials. In addition, 'old' responses for exemplars, $p = .052$, and new faces, $p = .142$, were comparable on one-face and two-face trials. In contrast, ‘old’ responses for averages were higher on two-face compared to one-face trials, $p < .001$.

Figure 2.2

Proportion of ‘old’ responses for exemplars, averages, and new faces for one-target and two-target trials.



Discussion

This experiment examined ensemble coding with an optimised paradigm, in which observers either encoded a single face or two faces. Ensemble coding was then assessed by testing recognition of the average of two encoded faces, exemplars of those faces, or an average of an encoded and a new face. In accordance with ensemble coding, the average was recognised less frequently when observers encoded a single target as opposed to two targets. The average, containing of 50% of each target, would have the same resemblance to a single target regardless of whether participants encoded one or two targets. Therefore, these initial findings likely reflect an internal representation that is an average, as the external average is more likely to be recognised when multiple targets are encoded, despite stable resemblance to the exemplars. Additionally, the average was selected as often as the exemplar in the two-face conditions, despite the fact that the former is not a direct visual match to either of the individual identities that had been encoded. Moreover, probe faces that were not encountered at encoding (new faces) were selected rarely, indicating that responses to averages were not false positives but reflective of the identities that had been encoded.

These findings provide initial evidence that observers are encoding an internal average. On the one hand, observers were more likely to select the average face when encoding two targets than they were if they encoded only one target. If the average was being selected based off its visual resemblance to an encoded exemplar, then selection of the average should be similar in these conditions, as the average would contain the same resemblance to a particular exemplar from the learning set, regardless of whether participants had encoded the second identity or not. This indicates that the average was selected because it resembles an internal representation that was formed by combining the learned exemplars. On the other hand, observers were as likely to select an average of two identities as they were to select an exemplar, suggesting that individual identities were also encoded. Whether the average can be encoded alongside individual exemplars, or whether selection of an average was increased because both encoded identities share resemblance to this, remains unclear.

Experiment 2

Experiment 1 provides initial evidence for ensemble coding, by indicating that the average is not recognised simply based on its resemblance to a single exemplar. However, the average and exemplar were also selected with similar frequency. In that experiment, the same image of each identity was used during the encoding and recognition phases. This makes it difficult to determine whether these effects reflect the encoding of identity or simpler image-based effects (Bruce et al., 1999; Burton, 2013; Hancock et al., 2000; Longmore et al., 2008; Megreya & Burton, 2006b). It is evident that identity encoding must be robust across changes in image properties for within-person variability to be tolerated (Jenkins et al., 2011), but evidence is mixed as to whether ensemble coding is as robust to these changes in images (e.g., Neumann et al., 2013; Rhodes et al., 2015; Sama et al., 2019). In a study by Davis et al. (2024), participants were able to recognise a face average across changes in viewpoint, but not across changes in image, implying ensemble coding effects are more likely to be image- than identity-based. To examine this possibility, Experiment 2 attempted to replicate the finding of Experiment 1 by using different images of the same identities at encoding and recognition. If the two-face average is selected at the recognition phase under these conditions, then this would provide evidence that ensemble coding of *identity* is occurring.

Methods

Participants

Fifty undergraduate students (11 male, 39 female) from the University of Kent, with a mean age of 20.4 years (*SD* = 5.0), took part in the experiment in exchange for course credit.

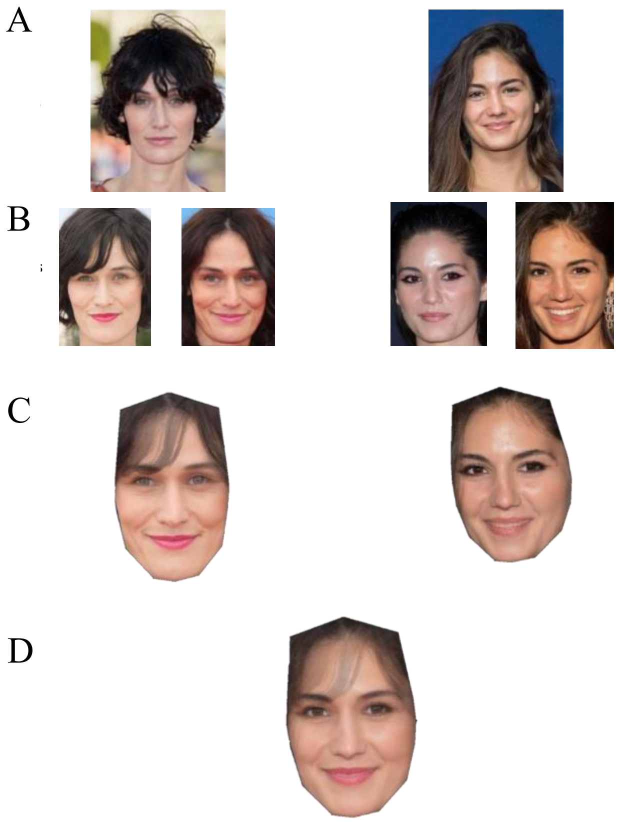
Stimuli and Procedure

The design and procedure were the same as in Experiment 1 except for the following changes. Three different images were employed for each identity. One of these images served as the image for the encoding phase. The other two images were then landmarked and averaged together to create the exemplar face. The two exemplars from each pairing were then averaged together to create the average as in Experiment 1. Exemplars and averages were used during the recognition phase, with exemplars who were unseen during the encoding phase acting as new identities.

As in Experiment 1, the experiment was programmed on PsychoPy (Version 2; Peirce et al., 2019) and conducted online through Pavlovia. On each trial, participants were either shown a single face or two faces during an initial encoding phase, followed by a recognition phase, and had to decide whether this identity had been seen previously (see Figure 2.3 for an example of stimuli).

Figure 2.3

An example of the stimuli used. Participants were shown an image of the identity during the encoding phase (A). Two different images of the same identity (B) were averaged together to create the exemplars (C). The exemplars were then averaged together to form the average (D).



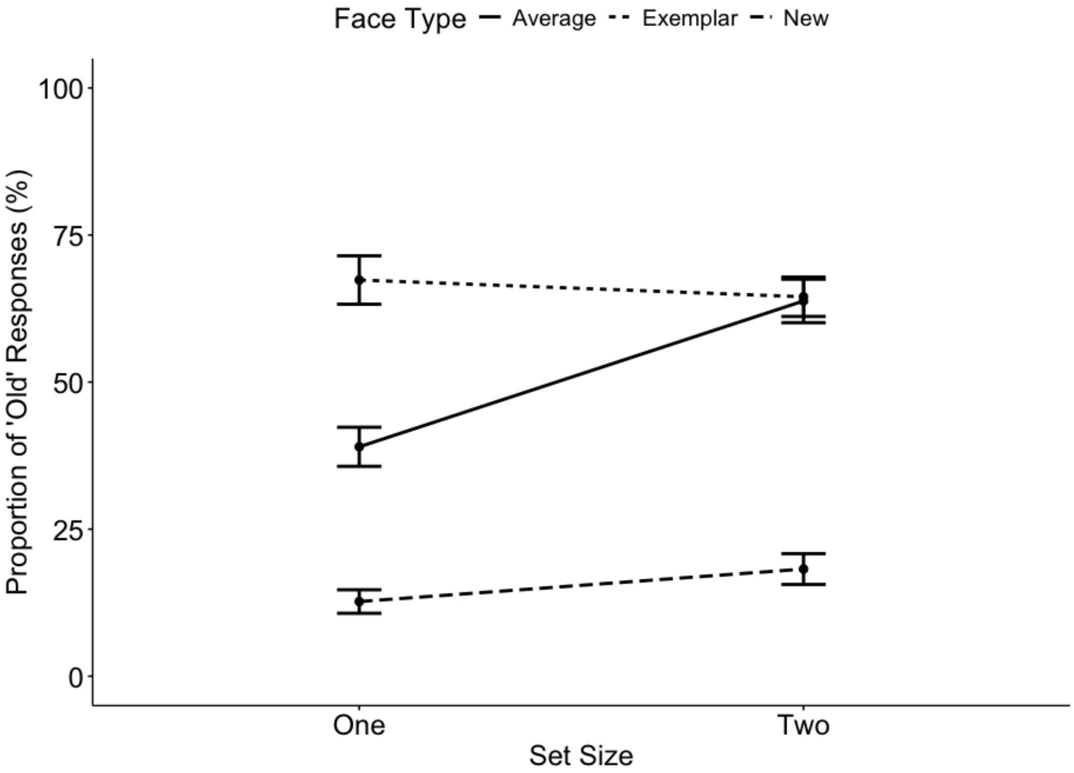
All participants complete 120 trials, comprising of 40 trials each for exemplar, average, and new face conditions. The experiment took approximately 45 minutes to complete. This was followed by the familiarity check for all identities.

Results

All trials on which an identity was familiar to participants were removed from analysis. This resulted in 206 trials being excluded, equating to 3.4% of all trials. ‘Old’ responses to the new faces were low, at 12.7% (*SD* = 10.9) and 18.2% (*SD* = 14.3) for one- and two-face trials, indicating the new faces were correctly rejected on the majority of trials. ‘Old’ responses for exemplar and average trials are shown in Figure 2.4. These data appear to replicate the pattern of responses in Experiment 1, with lower ‘old’ responses for the average compared to the exemplar in one-target trials but equal old responses for both conditions on two-target trials.

Figure 2.4

Proportion of ‘old’ responses to exemplar, average, and new faces for one- and two-target trials.



A 2 (Set Size: One vs. Two) x 3 (Face Type: Exemplar vs. Average vs. New) repeated-measures ANOVA of these data showed a main effect of set size, $F(1, 29) = 21.04$, $p < .001$, $\eta_p^2 = .42$, due to higher 'old' responses on two-face than one-face trials. There was also a main effect of face type, $F(2, 58) = 172.26$, $p < .001$, $\eta_p^2 = .86$, reflecting more 'old' responses for exemplar ($M = 65.92$, $SD = 20.40$) than new ($M = 15.45$, $SD = 12.94$), $p < .001$, and average faces ($M = 51.39$, $SD = 22.81$), $p < .001$, and more 'old' responses to average faces than new faces, $p < .001$. In addition, an interaction between these factors was found, $F(2, 58) = 22.84$, $p < .001$, $\eta_p^2 = .44$. Tukey's HSD showed that the proportion of 'old' responses was comparable across one- and two-face trials for new faces ($M_{one} = 12.69$, $SD_{one} = 10.94$; $M_{two} = 18.21$, $SD_{two} = 14.32$), $p = .143$, and exemplar faces ($M_{one} = 67.35$, $SD_{one} = 22.53$; $M_{two} = 64.50$, $SD_{two} = 18.30$), $p = .942$, whereas this was higher for the average in two-face ($M = 63.79$, $SD = 20.24$) than one-face ($M = 39.00$, $SD = 18.20$) trials, $p < .001$. Both exemplar and average faces resulted in more 'old' responses than new faces across one- and two-face trials, *p's* < .001. Most importantly, whereas 'old' responses for the exemplar were higher than for the average in one-face trials, $p < .001$, the percentage of 'old' responses for both conditions was comparable in two-face trials, $p > .999$.

Discussion

This experiment examined whether the pattern of results from Experiment 1 persist when different images of the same identities are used at encoding and recognition. This is an important test to determine whether any effects reflect face identity processing or simpler image-based effects (e.g., Davis et al., 2024). As in Experiment 1, averages were selected more frequently when observers encoded both targets compared to when they encoded a single target. Additionally, the number of old responses were comparable for exemplars and averages in the two-face condition, and much higher than for new faces that had not been seen during the initial encoding phase. This demonstrates that the average, which reflects a combination of identities that were seen at encoding, is selected as frequently as its constituent identities. This suggests that ensemble coding of the exemplars is occurring.

Experiment 3

The experiments so far indicate that averages are selected at a comparable rate to their constituent exemplars, indicating that these two face representations might have an equal status in observers' memory. This was observed when the same images were employed at encoding and recognition in Experiment 1 and across a change in image in Experiment 2, suggesting ensemble coding at the level of identity. However, the more frequent recognition of the average compared to the new face does not *necessarily* indicate that ensemble coding is occurring. Literature from the eyewitness domain demonstrates, for example, that in target-absent line-ups, observers will select the most similar identity to the target (Flowe & Ebbesen, 2007; Wells, 1993). However, for the same items, observers will select the matching target when this is present in the line-up. This corresponds to the face matching literature, in which misidentifications on mismatch trials do not correlate with match decisions to the same items on match trials (Fysh & Bindemann, 2023). These decisions are based on the visual resemblance between the compared faces, with errors on both match and mismatch trials correlating with the similarity ratings of the face to the target (Fysh & Bindemann, 2023). As yet, research has not put the average and the exemplar in direct competition with each other. Under this comparison, the exemplar should be *more* similar to the encoded identities than the average, because the exemplar is not contaminated by a second identity, unless participants are encoding an internal average.

In Experiment 3, the status of exemplars and averages were assessed directly, by contrasting the selection of both types of stimuli in the same recognition phase. For this purpose, the current paradigm was modified to present two test faces in a forced-choice paradigm. These test pairings comprised of a new and either an exemplar or an average face, or of an exemplar *and* an average. It is possible that averages are selected as often as exemplars when only a single probe face is presented (as in Experiments 1 and 2) or in competition with a previously unseen face. The question of main interest here is whether exemplars are selected more often than averages when these are in *direct* competition, by virtue of the exemplars' closer correspondence to the encoding set.

Methods

Participants

Sixty participants (9 male, 51 female) from the University of Kent, with a mean age of 19.6 years (*SD* = 3.2), took part in the experiment in exchange for course credit.

Stimuli and Procedure

The same stimuli as in Experiment 2 were employed. However, in this experiment the design was modified so that participants were always presented with two faces during encoding. Moreover, two faces were also presented during the recognition phase as opposed to a single face. For this, trials were counterbalanced so participants either saw pairings at recognition of either an exemplar and a new face, the average and a new face, or the exemplar and the average. Identities not used as an exemplar at recognition were used as new faces in later trials. This resulted in a total of 120 trials, with 40 trials per condition.

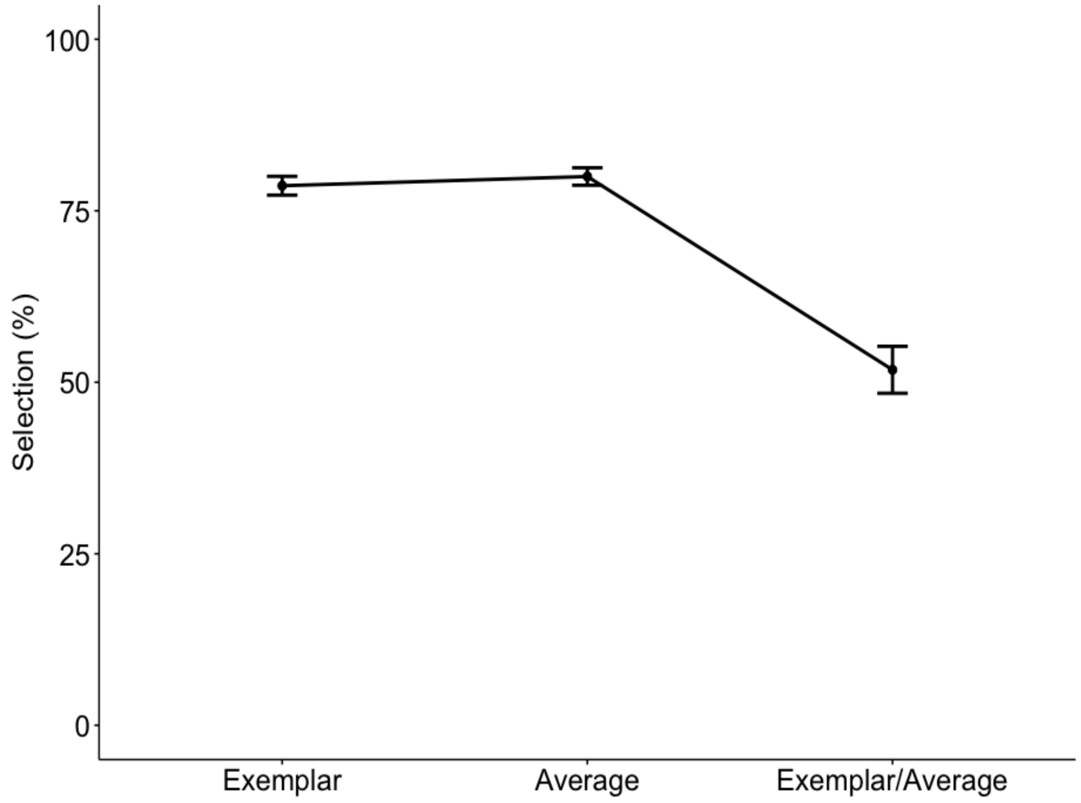
In each trial, a fixation cross appeared on screen for 1000 ms, followed by an encoding display of two faces for 4500 ms. This was followed by a second fixation cross for 1000 ms, and a recognition display comprising of a new face and either an exemplar or an average, or both an exemplar and an average. Participants were asked which of the two faces was one of the two targets they had learnt by pressing ‘S’ for the face on the left or ‘L’ for the face on the right side of the display. This was followed by the familiarity check for all identities.

Results

Any trials which included an identity familiar to a participant were excluded from analysis. This resulted in the removal of 3590, or 49.9%, of trials. The percentage of trials on which the exemplar and average were selected over the new face was then calculated and, for exemplar-average displays, the frequency with which the exemplar was chosen over the average. This data is shown in Figure 2.5. A one-factor ANOVA of these data showed an effect of condition, $F(2, 116) = 49.66$, $p < .001$, $\eta_p^2 = .46$. Tukey’s HSD test revealed that exemplar ($M = 78.65$, $SD = 10.61$) and average ($M = 80.00$, $SD = 9.88$) selections were comparable when these were displayed alongside a new face, $p = .471$, and were both higher in comparison with the exemplar-average pairings ($M = 48.18$, $SD = 26.28$), both p’s < .001. Considering that exemplars and averages were selected with similar probability when these stimuli were paired in the same display, performance across conditions was also compared to the midpoint of the accuracy scale (i.e., chance or 50%) via one-sample *t* tests. This showed that both exemplars, *t*(59) = 20.92, *p* < .001, *d* = 2.70, and average faces, *t*(59) = 23.52, *p* < .001, *d* = 3.04, were selected above chance level when paired with the new identity. In contrast, exemplar and average selections were at chance when these stimuli were paired, *t*(58) = 0.53, *p* = .597, *d* = 0.07.

Figure 2.5

Frequency that the exemplar and average faces were selected over the new face, and the frequency exemplars were selected over averages.



Note. The *y*-axis represents the number of times the face (exemplar or average) was chosen over the new face or the number of times the exemplar was chosen over the average.

Discussion

The aim of this experiment was to determine whether the average would be selected at a different rate to the exemplar when these were contrasted directly in the same display. If ensemble coding was occurring, it would be expected that the average would be selected more often than the exemplar. However, if the average was only selected in the previous experiments due to its resemblance to the exemplar, then when given both faces to choose between, the exemplar should be selected at a higher frequency. Neither was found to be the case, with the average and the exemplar being selected at an equal rate.

One potential explanation for this finding is that participants were only encoding one of the two identities. The average identity contains 50% of the information from each target, and, although the exemplar contains 100% information from one target, this would not provide any benefit over the average if this exemplar was not the identity participants had encoded. On the other hand, the average was not selected as frequently in Experiments 1 or 2 when participants only encoded one compared to two identities, suggesting the average is not selected based off resemblance to a single exemplar. This study therefore provides evidence that participants encode both exemplar and ensemble information equally.

General Discussion

This study examined the ensemble coding of identity under optimised conditions to determine the relative strength of exemplar and ensemble representations. Experiment 1 demonstrated that the average was recognised at a comparable rate to the exemplar when only two faces were encoded, corresponding with previous evidence for the ensemble coding of identity (Ji & Hayward, 2021;

Neumann et al., 2013; Neumann et al., 2018; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2018; Robson et al., 2018). This experiment used the same image at both encoding and recognition, which could have given the exemplar the advantage by providing a direct match to the image participants initially encoded. The average, on the other hand, was never viewed before.

To test the strength of the exemplar representation against that of the average under more comparable conditions, Experiment 2 repeated Experiment 1 using different images at encoding and recognition. This provides a stronger test for the coding of identity, as observers must determine whether two images show the same person rather than the image being identical (Bruce et al., 1999; Burton, 2013; Hancock et al., 2000; Longmore et al., 2008; Megreya & Burton, 2006b). The same pattern of effects was found, with comparable selection of the average to the exemplar. In contrast, previously unseen faces were rarely selected as probes. This pattern of results implies that both the exemplar and the average contain sufficient identity information for observers to match them to an encoded identity. However, the equivalence of these face categories in Experiment 1 and 2 also raised the question as to which face type participants would tend towards when they are required to make a direct comparison between the average and the exemplar.

Experiment 3 therefore investigated which face would be chosen when both were put into direct competition. To do this, participants were asked to encode two identities. This was followed by a recognition display of two probe faces and participants were asked to report which most resembled one of the two encoded identities. Again, exemplars and averages were selected at similar frequency. These findings therefore provide support for the idea of ensemble coding, in the sense that the exemplar image is a direct identity match to an encoded target and therefore should have a recognition advantage yet is not selected more frequently than the average.

As ensemble paradigms have not yet investigated ensemble coding with a single target face, the display times used in this study were based on previous literature on multiple perpetrator identifications (Bindemann, Sandford, et al., 2012). This was done to ensure sufficient time was given to encode both one- and two-target displays. However, most ensemble paradigms use display times of 2000 ms or less (e.g., Ji & Hayward, 2021; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). This study therefore provided participants with *more* time to encode the targets. Previous studies on ensemble coding suggest that with longer encoding durations, recognition of the exemplar increases (Li et al., 2016; Neumann et al., 2018). For emotion ensemble coding, recognition of the average also increases (Li et al., 2016). However, for identity, recognition of the average stable across durations varying between 50 ms and 6400 ms, with higher recognition scores for the exemplar than the average from 3600 ms (Neumann et al., 2018). The experiments within this chapter used a 4500 ms encoding duration for the two-targets, which could have given the exemplar an advantage over the average. Despite this, the experiments in this chapter found comparable identification of both the exemplar and the average.

These findings suggest that the average contains sufficient identity information to be classified as a match to the encoding images. This effect is strong enough to enable competition with the exemplar. At first glance, this appears to be consistent with other studies of ensemble coding (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019) and therefore provides converging evidence for the existence of this process in face identification. However, it is also possible that this effect arises due to the visual similarity between the encoding and the probe faces, rather than reflecting the formation of an internal cognitive representation that is based on an ensemble. Accordingly, the average face might be selected at recognition because it bears sufficiently resemblance to the learned faces even though, or because, it is a combination of the two. This similarity can sometimes be high – when participants have successfully encoded both encoding identities - but it can also be low if participants only encode one identity. However, and as Experiments 1 and 2 show, even such low similarity is sufficient for the average to be selected at least *some* of the time, on 30-40% of trials.

Similarly, selection of the exemplar at recognition should be best if observers have encoded *both* faces, as this provides the highest chance that the probe faces can be matched to one of the two learned faces. In contrast, if observers only encode one of the faces *and* the other identity is shown as the probe face, then performance would inevitably suffer. Crucially, such conditions could create an advantage for the average in the two-face conditions in comparison to the exemplar conditions: the average might be recognized based on its resemblance to either one of the exemplars, whereas the exemplar probe cannot be recognised if only the other exemplar has been learned. Across the many trials of an experiment, this could boost selection of the average relative to the exemplar, leading to comparable performance in both conditions.

Additionally, recognition of the average was higher in one-target trials than two-target trials (Experiments 1 and 2). This was taken as support for ensemble coding, as the average contains 50% of a single exemplar, regardless of whether participants encoded one or two targets. Therefore, if the average was being selected based on similarity, then it should be selected with comparable frequency across one-

and two-target trials. As this was not the case, this pointed to an internal average. However, it is also possible that in two-target trials, the average resulted in higher overall display similarity, in that the familiarity of the average to each encoded target was combined and facilitated recognition of this face.

This issue is complicated further as our study employed only two faces at encoding to examine ensemble coding in its most basic form, whereas most studies on the ensemble coding of identity use four or more faces. Under these conditions, the similarity of the average to each encoding face is reduced further. With two faces, for example, the average contains 50% of each identity. However, this is reduced to 25% with four faces. Previous research has shown that participants can still recognise an encoded face when it is averaged with a new identity, even if the second identity comprises only 30% of the average (Rotshtein et al., 2005). The experiments reported in this chapter are consistent with this finding by showing that participants will still identify a target when it is averaged with 50% of a second identity. However, Rotshtein et al. (2005) also demonstrate that if the average constitutes 70% of a new identity, it is no longer classified as a match to the encoded identity. This raises the question of whether observers can actually detect the resemblance of an exemplar to an average under the conditions that are typically applied in studies of ensemble coding, where each identity only constitutes 25% to an average (Bagaïni & Hole, 2017; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2013; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). It may be that the resemblance between learned exemplars and the averages faces shown at test is still strong enough for participants to detect, and therefore similarity may still be able to explain the equal recognition rates seen in these studies. However, if such low levels of resemblance cannot be detected by participants, then this supports the notion that an internal ensemble representation must have been formed by participants, which then triggers selection of these faces at test. This is examined further in Chapter 3.

CHAPTER 3:

Identification of the Constituent Identities of an Average

Introduction

Previous studies on the ensemble coding of identity demonstrate inconsistency in the strength of ensemble coding effects. Some provide evidence of more frequent recognition of an average compared to the exemplar (Matthews et al., 2018; Peng, Kuang, et al., 2019; Rhodes et al., 2015) and others show comparable recognition frequency (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019). However, these studies use complex designs, typically presenting participants with four or more identities to encode (e.g., Bagaïni & Hole, 2017; Matthews et al., 2018; Peng, Kuang, et al., 2019; Sama et al., 2019). Therefore, throughout Chapter 2, the ensemble coding of identity was investigated under optimised conditions, where two faces – the minimum needed for ensemble coding to occur – were presented at encoding. To further test for possible effects of resemblance of the average to an encoded exemplar, an additional condition was included where participants were given a single target to encode. Participants were then provided with a single recognition probe face, comprising of an exemplar, an average, or a new face. They were asked to determine if the face was old or new (Experiments 1 and 2) or to decide between two probe faces (Experiment 3).

With the same images used at encoding and recognition (Experiment 1), across a change in images (Experiment 2), and when put into direct competition (Experiment 3), the average was recognised at a similar frequency to the exemplar. The comparable selection of the exemplar and the average leaves it unclear as to what the internal representation of the encoded identities is. If this representation is an exemplar, then this should have higher similarity to an external exemplar, thus resulting in more frequent selection of this face compared to an average. The same logic would apply for the average, whereby an *internal* average representation should resemble an *external* stimulus more than the exemplar. However, as comparable selection was observed for both types of stimuli, this leaves open the possibility that the average is being recognised based off its resemblance to an encoded exemplar, rather than an encoded ensemble average.

However, across Experiments 1 and 2, the average was selected with higher frequency when participants encoded two targets than when they only encoded a single target. If the ensemble effects seen in the literature are occurring because observers are able to see the resemblance between a set average and the constituent exemplars, then participants should recognise the average with equal frequency across one- and two-target trials. This is because the resemblance of a single exemplar to the average will remain the same regardless of whether participants encoded the one target or both targets. However, across Experiments 1 and 2, the average was selected with higher frequency when participants encoded two targets. This provides initial support for the ensemble coding of identity as if the average was not selected because of resemblance to a specific exemplar, then it must be recognised based on its resemblance to the internal representation.

This argument is built on the assumption that observers can see the similarity between an average and its constituent exemplars. A parallel to the recognition of average and exemplar faces comes from face matching, whereby observers compare two images simultaneously to determine if the images depict the same or two different identities. Research from this field suggests that similarity is essential for unfamiliar face matching, whereby similarity between face pairs correlates with accuracy on match trials, in which two faces of the same person are shown, and mismatch trials, in which two face photos of different people are combined (Fysh & Bindemann, 2023). However, this task is often found to be difficult, with error rates of up to 45% (e.g., Fysh & Bindemann, 2018; Henderson et al., 2001). Although similarity must also be essential for accurate face recognition, it is unclear how adept observers are at spotting this between an average constructed of multiple faces and its constituent identities.

The extent of averaging that is accepted before the image is no longer considered a match has received some attention in previous work using facial morphs in matching paradigms. As less of the target identity is present within the morph (i.e., 30%), the less likely participants are to accept the morph as a match (Rotshtein et al., 2005). There is some variation around this, with some participants accepting up to 70% of the original target, and others accepting morphs with only 30% of the target (Hsu & Lee, 2016). However, an 80:20 morph is recognised as the identity constituting 80% of this as often as an image of that identity (Jenkins & Burton, 2011), which demonstrates that 20% of an identity can go undetected within a morph. In the majority of ensemble coding paradigms, four-face averages are used. This would result in a single encoded identity only contributing 25% to the average – less than the minimum amount found to be accepted by Hsu & Lee (2016) and not dissimilar to the 20% detection threshold (Jenkins & Burton, 2011).

However, in ensemble paradigms, this situation is complicated by the fact that participants are shown all four images that constitute the average. Therefore, although the average would only contain 25% of each contributing identity, participants have seen all four identities at encoding, and therefore would be familiar with 100% of the average's components. At the same time, if participants are only learning a subset of faces at encoding (i.e., three identities out of the four) – for example, due to short stimulus display times (i.e., 50 ms, as in Leib et al. (2014) and Neumann et al. (2018)), then the average would contain some amount (i.e., 25%) of an identity they had not encoded. Studies have manipulated the number of identities used to create a morph. These show that the less identities are used (i.e., 2), the more likely they are to be reported as a match to one of the original encoded images, but this decreases as more identities (i.e., 8 or 16) are added to this (Heyer et al., 2019). This is in contrast to previous work on ensemble coding, which suggests that ensemble coding occurs regardless of the number of targets in the display (Leib et al., 2014; Neumann et al., 2018; Peng, Kuang, et al., 2019). This implies that recognition of the average may draw on a different mechanism.

One way to investigate this is to test if people are able to discriminate which identities are used to create an average. Previous research has demonstrated that participants are able to name a famous celebrity that was used to create a 20-face average (Little et al., 2012). However, for unfamiliar faces, this task seems to be more difficult. For example, when asked to rate the similarity between an average and one of its constituent identities (Heyer et al., 2019), results show that, overall, participants rate the similarity to be low, and this is negatively related to the number of identities in the average. This implies that similarity is difficult to spot between an unfamiliar average and its constituent identities. However, this is dependent on the task. For instance, in matching tasks (e.g., Heyer et al., 2019), the average may not reach the similarity threshold required for an observer to make a match decision, but this is not informative as to whether observers can see *any* resemblance between an exemplar and average.

In this chapter, the ability to spot the resemblance between a four-face average and its constituent identities will therefore be examined to determine if this resemblance could potentially explain the equivalent recognition of the average and the exemplar in Chapter 2. If participants are able to spot the resemblance between a four-face average and its constituent identities, performance on these tasks should be high. However, if participants are unable to spot this resemblance, then this provides support for the ensemble coding of identity, as observers are not utilising similarity to the physical images, but rather to an encoded *internal* representation. The aim of Experiment 4 is to determine if ensemble coding can be explained by the resemblance of the average to each identity by asking participants to spot the similarity between the four-face average and its constituent identities. Subsequent experiments sought to clarify whether the results of Experiment 4 could be attributed to the average itself or the number of identities to compare and whether this is influenced by adding memory demands. Therefore, participants were shown either a four-face average (Experiment 4), a two-face average, or four individual images (Experiments 5-7). This was conducted as a matching task (Experiments 4 and 5) and with an additional memory component (Experiment 6). Memory was then further investigated by manipulating working memory load (Experiment 7).

Experiment 4

Ensemble coding is based on the assumption that people encode an internal average. However, the primary counterargument for ensemble coding is that observers are not encoding an internal average, but rather can detect enough similarity between the average to the encoded exemplar to facilitate recognition. As yet, there has been very little research on how much similarity an observer can detect between an average and an exemplar. Current research suggests that whilst people are largely capable of detecting a familiar face in an average (Little et al., 2012), this becomes more challenging when a face is unfamiliar (Heyer et al., 2019). This experiment examined whether participants aware of the task were able to spot the resemblance between a four-face average and its constituent identities to determine if resemblance to an average consisting of multiple identities is possible to detect. Participants were shown a four-face average surrounded by four unfamiliar selection identities and asked to select which of these, if any, they believed were used to create the average. These identities could be matching, or non-matching and participants were free to select as many as they thought were used. Crucially, this paradigm minimises any memory demands and eliminates the need to form an internal representation of the average. Instead, it provides a direct measure of whether participants can detect the resemblance between a four-face average and its constituent identities. In this manner, the experiment sought to explore whether similarity *per se* or ensemble coding can explain the ensemble effect seen in the literature (e.g., Leib et al., 2014; Neumann et al., 2018; Robson et al., 2018; Sama et al., 2019).

Methods

Participants

Forty participants were recruited for this experiment (7 male, 33 female). Participants were recruited from the University of Kent using a volunteer sample in exchange for 3 credits towards their course. Ages ranged from 18-26 years (*M* = 19.23, *SD* = 1.59).

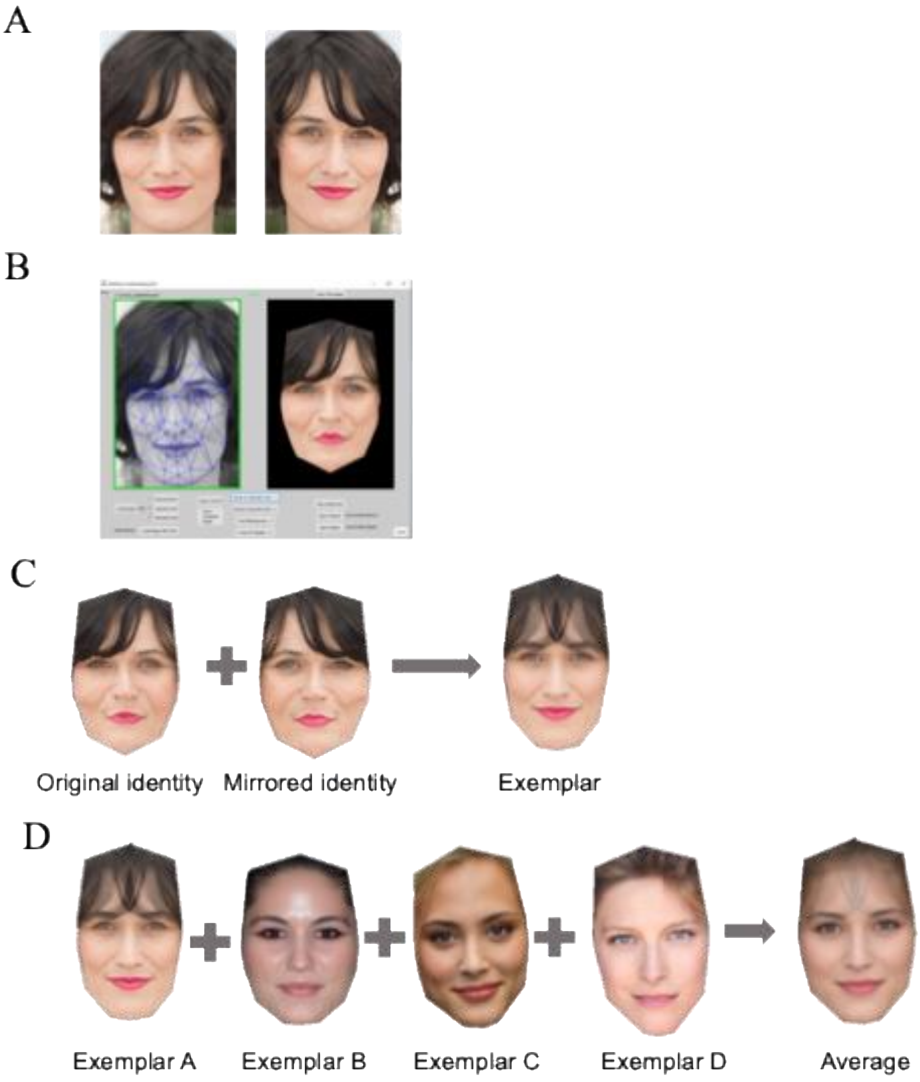
Materials and Design

The stimuli used for this experiment were the same as in Chapter 2. This consisted of three images of 240 French and German celebrities (120 male, 120 female) showing the face in a frontal view collected from a Google Image search.

One of these images were used as a possible selection identity, resulting in 240 selection identities. The remaining two images of each identity were used to create an exemplar. This involved using InterFace software (see Kramer, Jenkins, & Burton, 2017) to landmark key features of the face (eyes, nose, mouth, outline) using a polygon mesh which enabled the extraction of both shape (i.e., position of facial features) and texture information (lighting, reflectance) from each image. Texture information was then stored on a standard face template. Once extracted for both images, these were then averaged together, resulting in a single face that contains the average shape and texture of the two images – the exemplar (see Figure 3.1). This resulted in 240 exemplar faces. Identities were then randomly allocated to groups of four, with the restriction that identities were of the same sex and from the same country, resulting in 60 groups of four. The four exemplars in each group were then averaged together on InterFace to create the four-face average. Thus, overall, there were 60 average faces created.

Figure 3.1

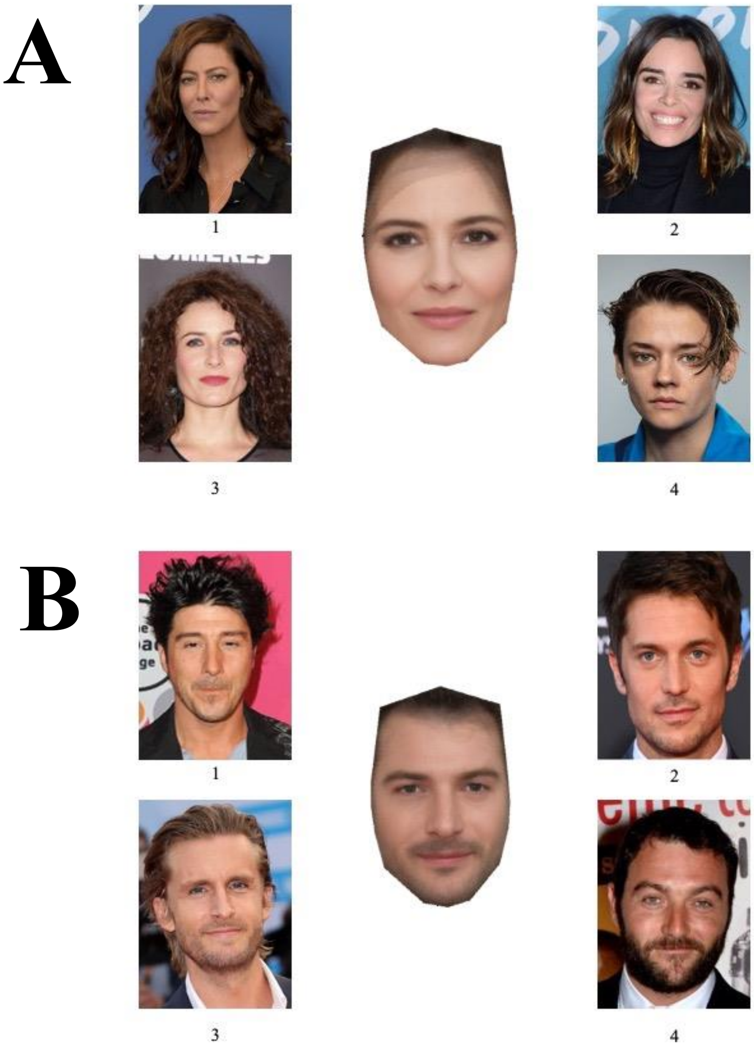
An illustration of how stimuli were created starting with the mirroring (A) and landmarking of the images (B) and then the averaging of the images to create the exemplar (C). Four exemplars were then combined in the same manner to create the cross-identity average (D).



On each trial of the study, the four-face average was displayed in the centre of the screen. Surrounding this, four selection images were displayed in each corner (see Figure 3.2). These selection images were either matching identities, new identities (identities taken from other groups), or a combination of the two. For each average face, selection displays with zero, one, two, three, and four matching identities were created, but these were counterbalanced across participants so that no identities were repeated. There were 60 trials in total comprising either no matching identities (12 trials) or one (12 trials), two (12 trials), three (12 trials), or four (12 trials) matching identities. Four-face averages were shown at a size of 118 (w) x 190 (h) pixels, with a resolution of 54 ppi. Selection images were shown at approximately 120 (w) x 163 (h) pixels, with a resolution of 54 ppi.

Figure 3.2

An example of the stimuli shown to participants. The four-face average in the centre surrounded by the four selection images. Identities 1, 2, and 3 (A) and identity 4 (B) were used to create the average.



Procedure

Participants were asked to complete a survey on Qualtrics which asked them for their demographic information as well as providing them with a consent form to fill out. Once this was completed, participants were allocated to one of five counterbalanced versions of the study. The study was programmed on PsychoPy

(Version 2; Peirce et al., 2019). Each trial featured the average in the centre of the screen, surrounded by four selection identities. They were asked to identify which of the four identities were used to create the average. To select a face, participants were asked to left click with the mouse on any faces they believed were used. Once a participant had selected a face, a semi-transparent green square appeared around that face to show the participant which faces they had selected. If participants wished to unselect a face, they were told to click again with the left mouse on the selected face which would unselect the face and the green square would be removed. Once participants were happy with their selection, or if they believed none of the four identities to have been used to create the average, they clicked on a ‘next’ button at the bottom of the screen. There were 60 trials in total with a break halfway through. For each identity, there were four possible outcomes. Participants could correctly identify the face as a constituent identity (a hit). However, if the selected identity was a new face (i.e., not used in the average), this would be classified as a misidentification. If participants did not select a face, this was either a correct rejection or a miss depending on if the face was new or a constituent identity of the average. For each trial, participants could score four different responses.

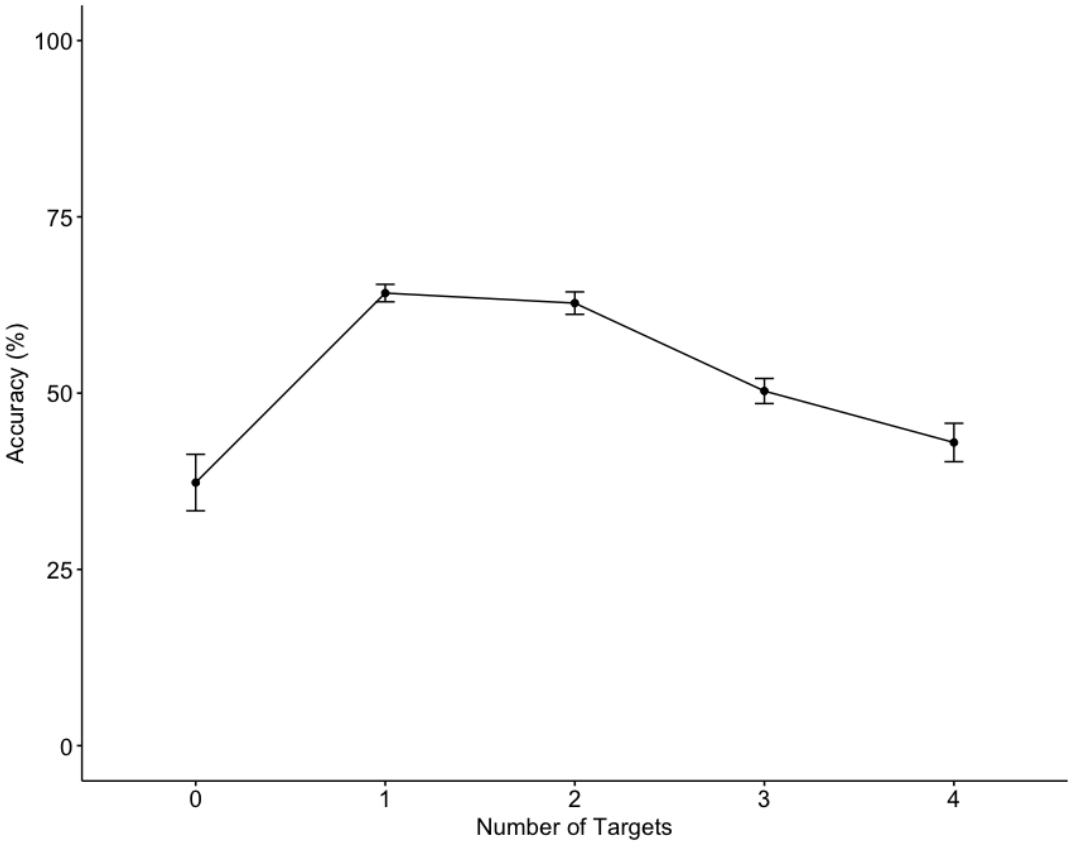
Following this, participants were presented with a familiarity check in which they were shown an image of each identity used in the study sequentially alongside four possible names numbered 1-4 and asked to press ‘S’ if they did not recognise the person, ‘L’ if they knew the person but not their name, or the number on the keyboard corresponding to the name of the identity. The experiment lasted approximately 45 minutes in total.

Results

Data was first cleaned by removing any trials that included an identity the participants rated as familiar, resulting in the removal of 538 trials (22.42% of all trials, *Mode* = 0%, *M* = 13.45%, *SD* = 15.47%). The percentage accuracy of participants' responses was then analysed combining hits and correct rejections. Overall, accuracy was at 51.6%. Mean accuracy reduced as the number of targets present in the display increased (see Figure 3.3). Accuracy was lowest with no targets in the display.

Figure 3.3

Mean accuracy per number of matching selection identities present in the display. Error bars show the standard error.



A one-way repeated-measures ANOVA of overall accuracy revealed an effect of target presence, $F(4, 148) = 19.34, p < .001, \eta_p^2 =.34$. Tukey's HSD test demonstrates that this arises from a decline in accuracy when more targets are present in a display, whereby accuracy was lower with three (*M* = 50.30, *SD* = 11.26) and four (*M* = 42.50, *SD* = 16.95) targets than in the one (*M* = 65.17, *SD* = 8.02) and two (*M* = 62.77, *SD* = 9.82) target conditions, all *ps* < .001. In contrast, accuracy was more comparable for one and two target displays, *p* = .852, and for three and four target displays, *p* = .058. Finally, accuracy for one and two targets was also higher than when no target was present in a display at all (*M* = 37.31, *SD* = 25.34), *ps* = .001. None of the other comparisons were significant, all *p*s > .181.

Hits

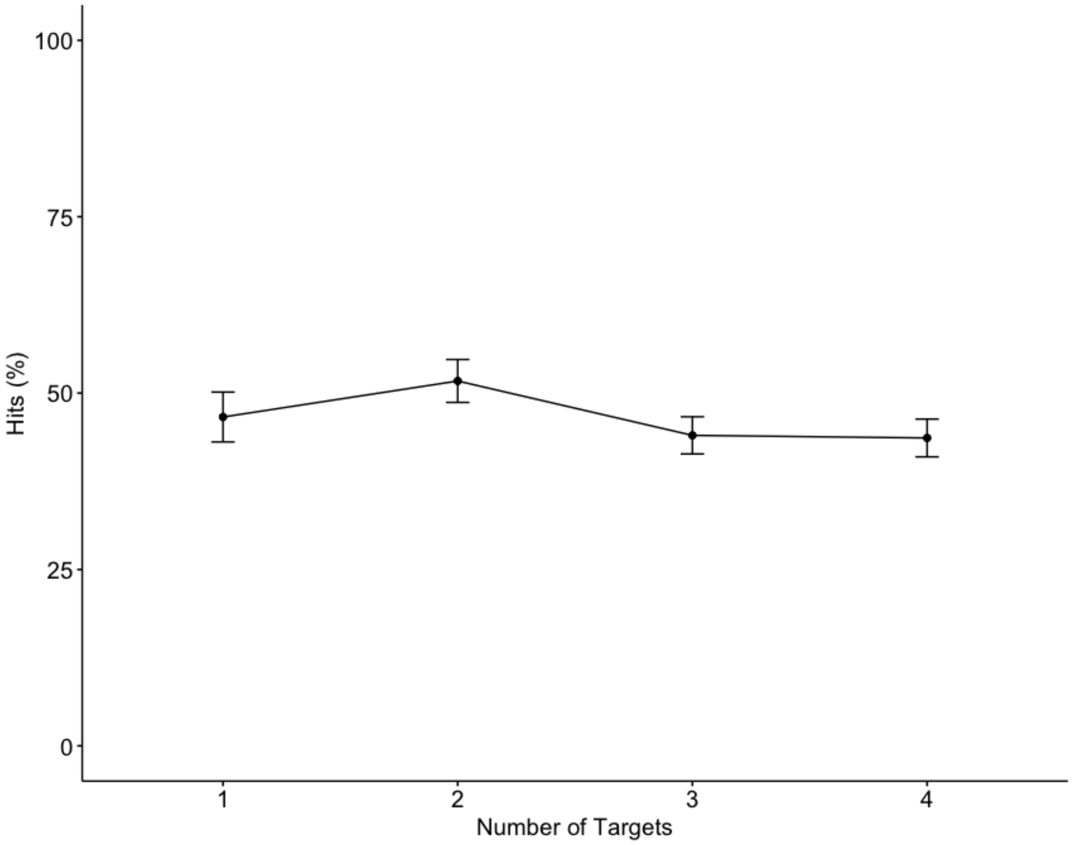
To analyse the data further, accuracy was split into hits, which corresponds to the correct selection of an exemplar, and correct rejections, which corresponds to correct non-matching decisions. The mean number of hits for each condition can be seen in Figure 3.4. A one-way repeated-measures ANOVA of this data showed an effect of target presence, $F(3, 111) = 4.74, p = .004, \eta_p^2 = .11$, as accuracy was higher with two than three and four targets, both *ps* < .011 (see Table 3.1). None of the other comparisons were significant, all *p*s > .413.

Table 3.1

	Hits			
	1	2	3	4
Mean	47.68	51.72	44.01	43.71
SD	22.92	18.75	16.66	17.03

Figure 3.4

Mean hit rate by number of matching selection identities present in the display. Error bars show the standard error.



Correct Rejections

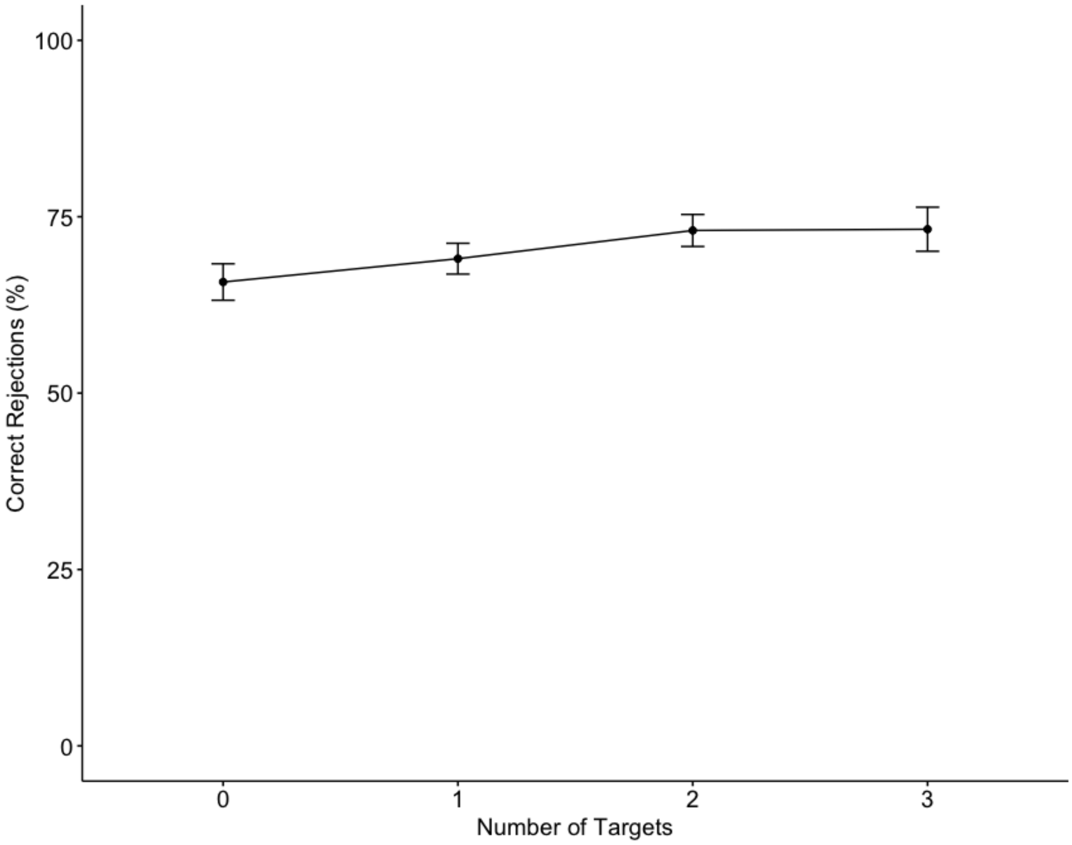
The proportion of correct rejections was lowest in the no target condition and increased up to the three-target condition (see Figure 3.5). A one-way repeated-measures ANOVA showed an effect of target number, $F(3, 108) = 3.63$, $p = .015$, η_p^2 =.09. Tukey's post-hoc tests revealed fewer correct rejections were made in the no target conditions than the two target condition, $p = .048$ (see Table 3.2). None of the other comparisons were significant, all *p*s > .083.

Table 3.2

	Correct Rejections			
	0	1	2	3
Mean	65.75	69.97	73.05	73.23
SD	15.93	13.76	14	19.76

Figure 3.5

Proportion of correct rejections by number of matching selection identities in the display. Error bars show the standard error.



Discussion

This experiment sought to investigate whether observers can spot resemblance between an average and its constituent identities to determine if observers can detect the similarity between a single identity and an average consisting of multiple identities. In this experiment, participants were presented with the average alongside the selection identities. There was no time limit, and so participants were free to make as many comparisons as needed to determine which identities most resembled the average, and memory demands were minimised. Despite this, overall accuracy was at approximately 50%. At first glance, this appears comparable to previous studies on the ensemble coding of identity, where the average is reported as a set member around 50-60% of the time (e.g., Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng et al., 2021; Rhodes et al., 2018). However, in these studies, the formation of an internal average is *assumed*. Here, participants were provided with the average to directly investigate whether the average could have been recognised based on its resemblance to encoded exemplars rather than the formation of an internal average. If the ensemble effects seen in previous research occurred due to the average's resemblance to the exemplars, then participants would be expected to perform well at this task – especially since this task did not require the *internal* encoding of the average but merely a direct visual comparison with its constituent identities. However, results showed that half of the identities used to create the average were missed by participants and non-matching identities were accepted as a match over 25% of the time. This demonstrates that participants had difficulty seeing the resemblance between a four-face average and its constituent identities under optimal conditions.

There was some evidence that participants could identify the faces that were used to create the averages. Generally, however, this did not depend on the number of correct targets in the display. Moreover, this likely interacted with participants' bias to select a certain number of faces. Observers tend to have a base-rate criterion, which reflects how many items they typically expect to be in an experimental display (e.g., Aminoff et al., 2012; Kantner & Lindsay, 2012). In this experiment, when no correct targets were present, correct rejection scores were lowest, suggesting a potential bias to select a face even when no correct targets were present. Additionally, the proportion of hits reduced as the number of correct targets increased, suggesting that this bias was for one or two faces.

Nevertheless, the results from this study demonstrate that observers have difficulty seeing the resemblance of an average to its constituent exemplars even when direct visual comparison is possible. This provides support for the ensemble coding of identity, as previous research shows averages are selected as often as exemplars (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019), as supported in Chapter 2. This study demonstrates that resemblance of an average to an exemplar is an unlikely explanation for this, and suggests that the average is being recognised because it matches what has been encoded internally – an ensemble.

Experiment 5

Experiment 4 demonstrated that detecting resemblance between a four-face average and its constituent identities is a difficult task. This relates to previous research, which has shown that observers fail to notice when an identity has been morphed with 20% of a new face (e.g., Hsu & Lee, 2016; Jenkins & Burton, 2011). Whether the difficulty in detecting the constituent identities of a four-face average arises because the average does not resemble the constituent identities or because participants have difficulty in processing the amount of identity information is unclear. Literature on the multiple-perpetrator effect and on dual-target visual search demonstrates that having to divide attention between targets impairs identification performance (e.g., Megreya & Bindemann, 2012; Mestry et al., 2017). This is further exacerbated by having to make additional comparisons between each target and the selection identities (Megreya & Bindemann, 2012). It is therefore possible that this inability to detect the resemblance may arise from having to compare four selection identities to the average.

Therefore, this experiment will include two further conditions – a four exemplar image condition and a two-face average condition. In the four exemplar image condition, four original images of the identities will be displayed in place of the average. This will require more comparisons than a single average face. If performance is reduced in these trials compared to the average trials, then difficulty in the task can be attributed to the number of comparisons. However, if the amount of identity information in the average can better explain the inability to detect resemblance, then the two-face average, which contains more of each constituent identity, should outperform the four-face average.

Methods

Participants

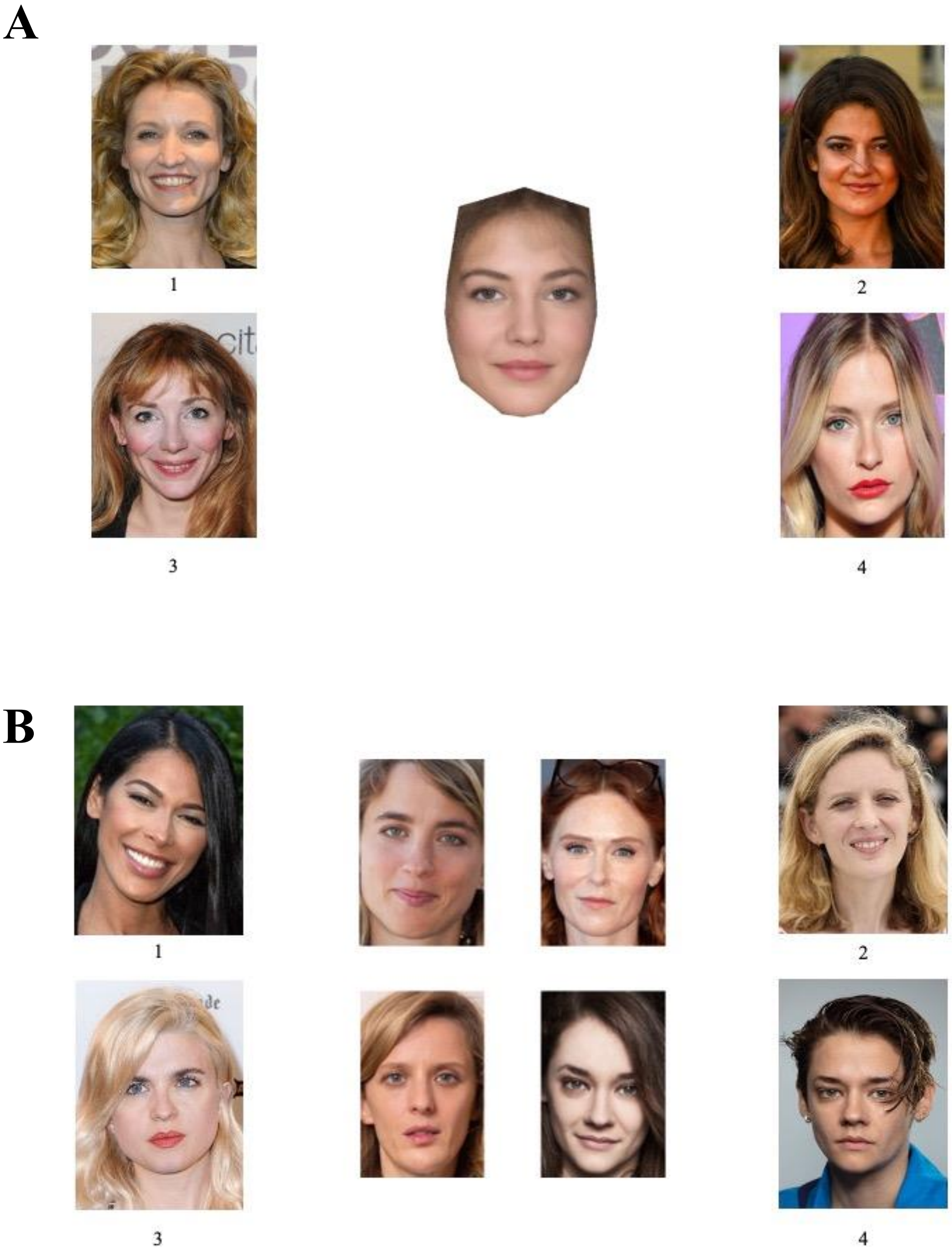
Thirty participants (3 male, 27 female) were recruited from the University of Kent in exchange for 3 credits towards their course. Ages ranged from 18 to 34 years (*M* = 20.07, *SD* = 3.08). To take part in this study, participants must not have taken part in Experiment 4.

Materials, Design, and Procedure

The paradigm was based on Experiment 4 using the same stimuli, with the addition of two more conditions: one which used a two-face average and another which presented the four targets individually (see Figure 3.6). To reduce the complexity of the design, instead of having 0-4 possible correct targets, each trial either had 50% of the number of faces in the set or 100%. For example, for trials in which the target was a two-face average either one of the constituent identities (50%) or both identities (100%) would be present in the four possible selection identities. For the four-face average and the four exemplar image conditions, either two out of the four would be present (50% condition) or all four would be present (100% condition). Participants could make a hit, miss, misidentification, false positive, or correct rejection, and could score up to four responses correct in one trial. Participants were told that targets may or may not be present, and the number of correct targets could be anything from one to all four of the identities. They were also informed that the selection identities could be a match to any of the four exemplars in the four exemplar image condition, regardless of their position on screen. Two- and four-face averages were displayed at 118 (w) x 190 (h) with a resolution of 54 ppi. Each target in the four exemplar image condition was displayed at 82 (w) x 133 (h) with a resolution of 54 ppi. Selection identities were shown at 120 (w) x 163 (h) with a resolution of 54 ppi. There were 48 trials with a break halfway through. Of these 48 trials, 16 were four-face averages, 16 were two-face averages, and the remaining 16 were four exemplar images. Each of these conditions were also split equally into 50% and 100% versions. Afterwards, participants were asked to complete the same familiarity check as Experiment 4.

Figure 3.6

An example of a two-face average (A) and of the four exemplar image (B) condition. Four-face average stimuli can be seen in Figure 3.2.

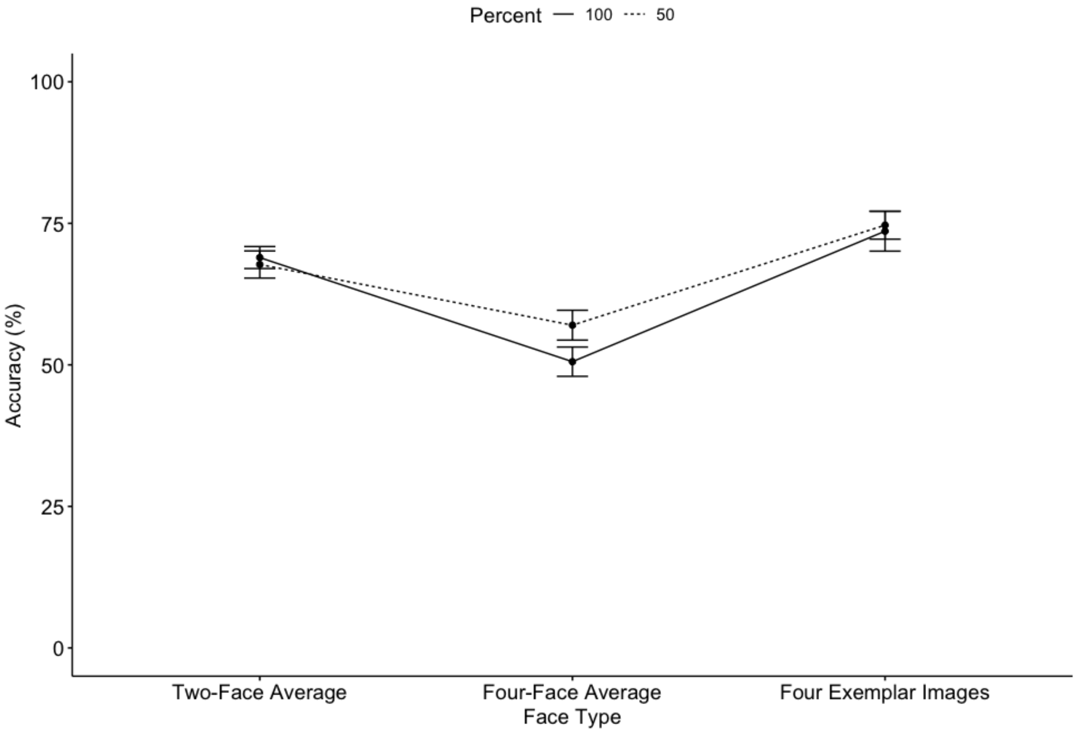


Results

The percentage accuracy was analysed after removing trials that included familiar identities. This resulted in the removal of 367 trials (25.49 % of all trials, *Mode* = 0%, *M* = 12.23%, *SD* = 9.88%). Overall accuracy, including both hits and correct rejections, was at 65.8% (see Figure 3.7). Accuracy was higher for the four exemplar images and the two-face average condition, with lowest accuracy for the four-face average conditions.

Figure 3.7

Accuracy across trial types for both 50% and 100% conditions. Error bars show the standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar images) x 2 (Target Presence: 50% vs. 100%) repeated-measures ANOVA was conducted on the data. No main effect of target presence was found, *F*(1, 28) = 2.10, *p* = .159, η_p^2 = .07, demonstrating that having 50% (*M* = 66.47, *SD* = 15.40) or 100% (*M* = 64.27, *SD* = 17.85) of the target identities did not affect the ability to detect resemblance.

However, there was a main effect of face type, $F(2, 56) = 32.90$, $p < .001$, $\eta_p^2 = .54$, whereby Tukey's HSD showed that accuracy was lower for the four-face average ($M = 53.79$, $SD = 14.55$) than either the two-face average ($M = 68.34$, $SD = 11.87$) or four exemplar images ($M = 74.15$, $SD = 16.29$), both *ps* < .001. In contrast, accuracy was comparable between the two-face average and four exemplar image trials, $p = .071$.

An interaction between these factors was also found, $F(2, 56) = 3.57$, $p = .035$, $\eta_p^2 = .11$. Tukey's HSD revealed a drop in accuracy for the four-face average when 100% of the targets were present compared to when 50% were present, $p = .027$ (see Table 3.3). Lower accuracy was also observed for four-face averages than either two-face averages or four exemplar images when 100% of targets were present, *p*s < .001. However, when 50% of targets were present, four-face averages only produced lower accuracy than the four exemplar images, $p < .001$. No other comparisons were significant, *p*s > .141.

Table 3.3

	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
Mean	67.73	68.95	57.01	50.56	74.67	73.61
SD	13.13	10.65	14.44	14.17	13.51	18.98

Hits

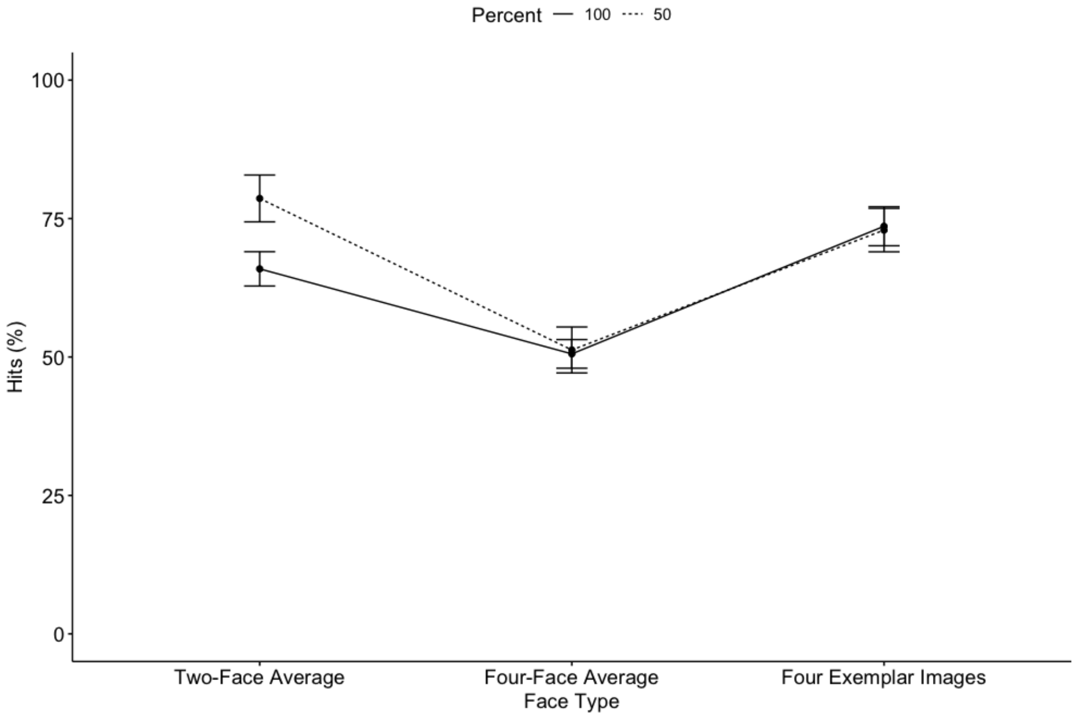
To further analyse the data, accuracy was divided into hits and correct rejections. Descriptive statistics for both hits and correct rejections can be seen in Figure 3.8. A 3 (Face Type: two-face average, four-face average, four exemplar images) x 2 (Target Presence: 50% vs. 100%) repeated-measures ANOVA was conducted on the data, showing a main effect of face type, $F(2, 56) = 30.83$, $p < .001$, $\eta_p^2 = .52$, as higher hits were observed for two-face averages ($M = 72.28$, $SD = 21.13$) and four exemplar images ($M = 73.26$, $SD = 20.12$) than four-face averages ($M = 50.92$, $SD = 18.79$), *p*s < .001, but this was comparable for four exemplar images and two-face averages, $p = .769$. There was also a main effect of target presence, $F(1, 28) = 10.04$, $p = .004$, $\eta_p^2 = .26$, with greater accuracy on 50% trials ($M = 67.61$, $SD = 25.18$) than 100% trials ($M = 63.25$, $SD = 19.20$), $p < .001$. The interaction did not reach significance, $F(2, 56) = 2.70$, $p = .076$, $\eta_p^2 = .09$ (see Table 3.4).

Table 3.4

	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
Mean	78.63	65.92	51.27	50.56	72.92	73.61
SD	23.16	16.98	22.74	14.17	21.49	18.98

Figure 3.8

Hit accuracy across face types by percentage of matching targets present. Error bars show standard error.

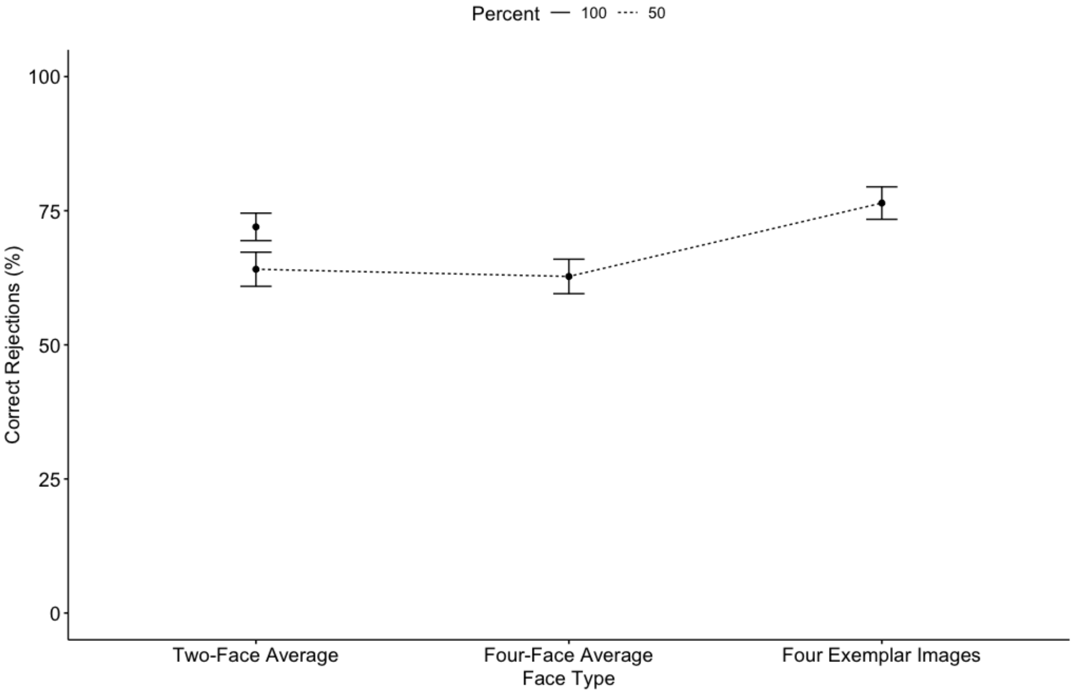


Correct Rejections

Correct rejections were lowest for both average conditions (two-face and four-face), but higher in the 50% condition for four exemplar image displays (see Figure 3.9). As correct rejections were not possible in the 100% of four-face ensembles or four exemplar image trials, a one-way repeated-measures ANOVA was then conducted on the 50% trial conditions. This demonstrated an effect of face type, *F*(2, 58) = 6.06, *p* = .004, η_p^2 = .17. Tukey's HSD analyses revealed higher correct rejections for four exemplar images (*M* = 76.43, *SD* = 16.58) than either four-face averages (*M* = 62.75, *SD* = 17.60), *p* = .024 or two-face averages (*M* = 64.09, *SD* = 17.42), *p* = .027. No difference was observed between two- and four-face averages, *p* = .922. Finally, a paired samples t-test was used to compare 50% to 100% trials for the two-face average. This revealed higher correct rejections on the 100% trials (both faces present in the four selection faces) (*M* = 71.99, *SD* = 14.02) than the 50% trials (one of the two faces present), *t*(29) = 2.47, *p* = .020, *d* = -0.45.

Figure 3.9

Graph showing correct rejection rate across face types by percentage of matching targets present. In the 100% of four-face average and four exemplar image trials, all faces were matches, and so no correct rejections are reported. Error bars show standard error.



Discussion

This experiment sought to investigate why resemblance between a four-face average and its constituent identities is difficult to detect. Previous research has demonstrated that having to make multiple comparisons impairs matching performance (Heyer et al., 2019; Megreya & Bindemann, 2012). In Experiment 4, observers had to compare four selection identities to a target average. To investigate whether the number of comparisons could have impaired performance, this experiment included a four exemplar images condition, which required each selection image to be compared to all four target images, resulting in four times the number of comparisons than the average condition. However, task performance was worse for the four-face average than the four exemplar images conditions, suggesting that low performance on this task is not attributable to the number of comparisons.

It is plausible that observers are unable to see the resemblance between a four-face average and its constituent identities in the first place. To test this, a two-face average condition was included. In these displays, each constituent identity contributes to 50% of this average, which has been shown to be sufficient for recognition (e.g., Kramer et al., 2019; Nightingale et al., 2021; Robertson et al., 2017). Performance for this condition was better than for the four-face average, and comparable to the four exemplar images. This suggests that the four-face average does not contain sufficient information about each identity for exemplar identifications to be made with greater accuracy. Therefore, this indicates that the ensemble coding effects seen in the literature cannot be explained as such by the resemblance of the average to the encoding images.

However, ensemble coding is a memory phenomenon, whereby observers have to internally *encode* the identities. Independent impairments can be seen in face matching and memory tasks, which suggests that these tasks recruit different cognitive processes (Stantić et al., 2022; Stantić et al., 2023). The average face might represent what is encoded internally and this may operate differently to the identity matching of exemplars to an external average – as was the case in Experiment 4 and 5. Therefore, it remains unclear whether these findings extend to a scenario in which observers have to determine resemblance to the exemplar based on the internal representation that is encoded (i.e., the average). This is examined in the next experiment.

Experiment 6

Experiments 4 and 5 demonstrate that people have difficulty in spotting the resemblance between a four-face average and its constituent identities, and that this likely reflects the lower similarity of an average to each of its identities (Experiment 5). However, memory is an essential component of ensemble coding. If participants encode an internal average, it is possible that the exemplar can then be recognised based on the resemblance to the encoded average. Memory for faces seems to rely on different mechanisms to face matching (Stantić et al., 2022; Stantić et al., 2023), and it is evident that an internalised representation must be capable of accommodating some level of variability in a person's appearance – otherwise only some face images could be identified but not others (Corpuz & Oriet, 2022; Kramer, Jenkins, Young, et al., 2017; Ritchie et al., 2017). Consequently, it is possible that what participants internalise during ensemble coding may provide an advantage when detecting resemblance to its constituent identities. This study therefore investigates participants' ability to identify the constituent identities of an encoded average (two- or four-face) or four exemplar images with a memory component.

In previous studies in this domain, participants are often shown a set of exemplars at encoding and the formation of an average of these identities is *assumed* (e.g., Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Neumann et al., 2018; Peng et al., 2022). Moreover, in such paradigms, the parallel encoding of exemplars cannot be ruled out. To circumvent these problems and ensure that an average is encoded that represents an even mixture of its constituent identities but not the exemplars, participants were asked to encode the average images directly. In this way, the internal representations of two- and four-face averages was controlled. As in Experiment 5, this was then contrasted with a condition in which four exemplar images were learned. In this paradigm, if participants' ability to spot the resemblance between a four-face average and the constituent identities matches their accuracy for four exemplar images, this would suggest that the four exemplar images are being encoded into an average and provide evidence that the internal representation is, in fact, an average.

Methods

Participants

Twenty participants were recruited using volunteer sampling at the University of Kent in exchange for 3 credits towards their course (1 male, 19 female). Ages ranged from 18 to 21 years (*M* = 19.20, *SD* = 0.62). To take part in this experiment, participants must not have completed either Experiments 4 or 5.

Materials, Design, and Procedure

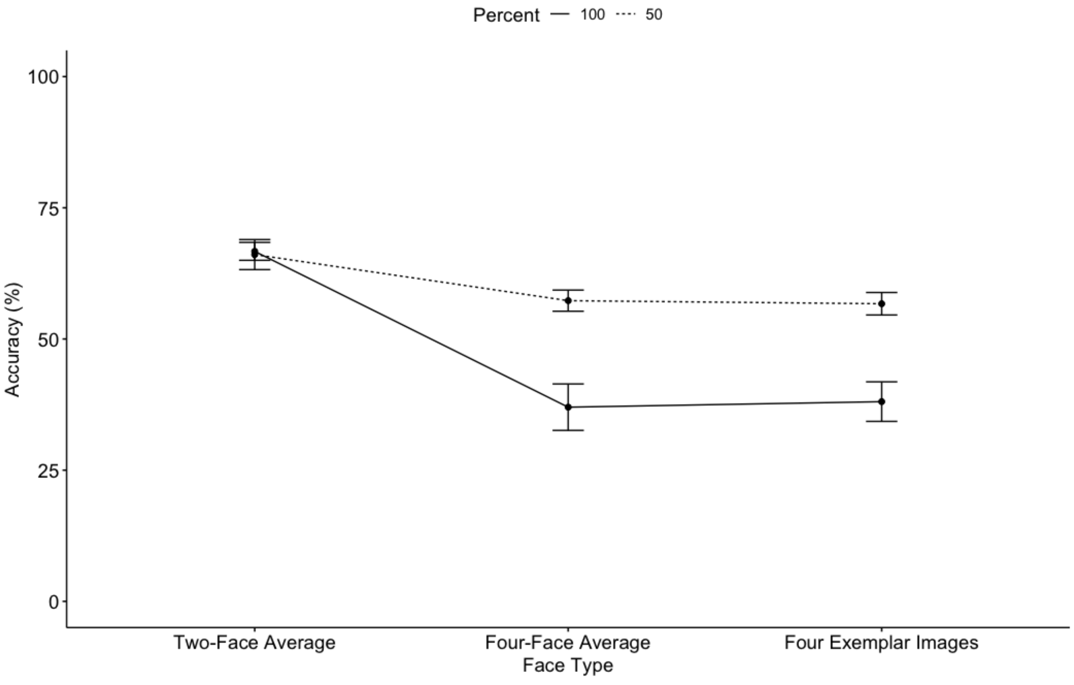
This study followed the same design and procedure as Experiment 5, however, for each trial, the target faces (either the two-face average, four-face average, or four exemplar images) were presented first without the selection identities in an encoding phase. These faces remained on screen for 4.5 seconds before being replaced by a fixation cross (1 second). The four selection identities were then displayed, and participants were asked to select which, if any, they believed were present in the encoding phase. There were 48 trials in total with a break halfway. Participants were also required to complete the familiarity check. The experiment lasted approximately 45 minutes.

Results

Firstly, trials including a familiar face were removed, resulting in the removal of 312 trials (32.50% of all trials, *Mode* = 16.00%, *M* = 15.6%, *SD* = 12.66%). Percentage accuracy was then calculated (see Figure 3.10). Overall accuracy was 52.8%. Accuracy was highest for the two-face average condition, across both 50% and 100% levels. Accuracy for the four-face average and four exemplar images followed a similar pattern, with a decline in accuracy in the 100% compared to the 50% conditions.

Figure 3.10

Accuracy for each face type by percentage of matching targets present. Error bars show standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar image) x 2 (Target Presence: 50% vs. 100%) repeated-measures ANOVA was conducted, which revealed a main effect of Face Type, *F*(2, 32) = 29.86, $p < .001$, $\eta_p^2 = .65$. Tukey's HSD demonstrated that both the four-face average (*M* = 46.88, *SD* = 18.06) and four exemplar image (*M* = 47.15, *SD* = 16.51) conditions resulted in comparable accuracy, $p = .931$, but both resulted in lower accuracy than the two-face average condition (*M* = 66.38, *SD* = 10.42), *p*s < .001. There was also a main effect of Target Presence, *F*(1, 16) = 25.73, $p < .001$, $\eta_p^2 = .62$, with higher overall accuracy in the 50% trials (*M* = 60.19, *SD* = 11.18) than the 100% trials (*M* = 46.75, *SD* = 20.47), $p < .001$. This was qualified by an interaction, *F*(2, 32) = 23.89, $p < .001$, $\eta_p^2 = .60$ (see Table 3.5). In the 50% condition, there were no differences between any of the three face type conditions, all *p*s > .126. Whilst accuracy for two-face average trials remained consistent across 50% and 100% conditions, $p = .992$, drops in accuracy were observed for both four-face average, $p = .009$, and four exemplar image trials, $p < .001$, when 100% of targets were present compared to 50%. This resulted in lower accuracy for the four-face average and four exemplar image trials compared to two-face average trials, *p*s < .001 with 100% targets present, but with no difference between four-face average and four exemplar image trials, $p = .973$.

Table 3.5

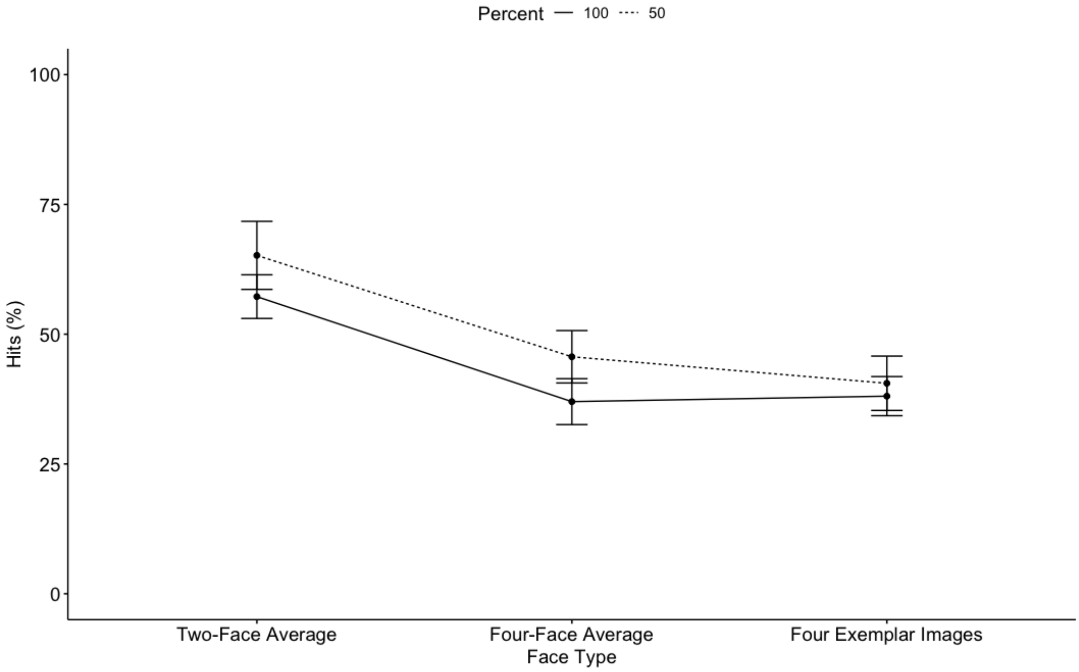
	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
Mean	66.08	66.7	57.3	37	56.72	38.06
SD	12.8	7.27	8.54	19.28	9.35	16.85

Hits

Data was then split into hits and correct rejections. Mean hits can be seen in Figure 3.11. Hits showed a similar pattern to the overall accuracy, with drops in the 100% compared to the 50% conditions, although for hits this was also observed in two-face average trials. Moreover, two-face average trials had higher hits overall than either four-face average or four exemplar images, which both demonstrated comparable hits.

Figure 3.11

Hit accuracy for each face type by percentage of matching targets present. Error bars show standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar images) x 2 (Target Presence: 50% vs. 100%) repeated-measures ANOVA demonstrated an effect of face type, $F(2, 32) = 25.89$, $p < .001$, $\eta_p^2 = .62$. Tukey's HSD analysis showed higher hits for the two-face average (*M* = 61.42, *SD* = 24.57) than either the four-face average (*M* = 41.21, *SD* = 20.53) or four exemplar images (*M* = 39.27, *SD* = 19.76), *p*s < .001, but no difference between the four-face average and four exemplar images, *p* = .398. In addition, a main effect of target presence was also found, *F*(1, 16) = 9.00, *p* = .008, η_p^2 = .36, hits were higher for 50% (*M* = 50.80, *SD* = 26.77) trials than 100% (*M* = 43.76, *SD* = 19.95) trials. The interaction between factors did not reach significance, *F*(2, 32) = 0.98, *p* = .388, η_p^2 = .06 (see Table 3.6).

Table 3.6

	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
Mean	65.18	57.24	45.64	37	40.54	38.06
SD	29.31	17.84	21.41	19.28	22.83	16.85

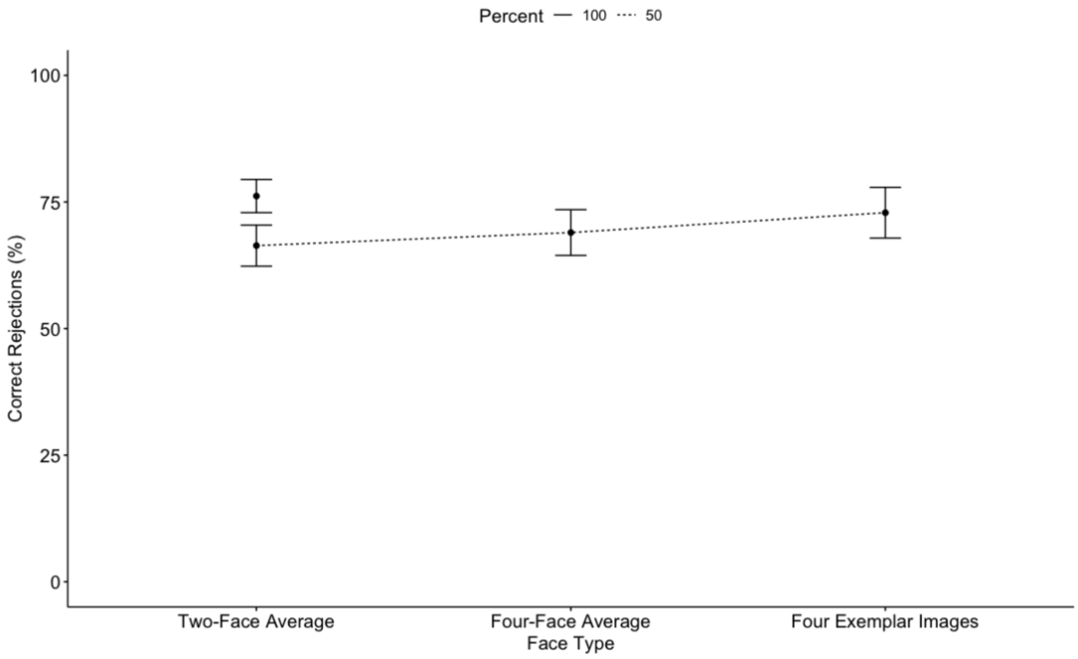
Correct Rejections

Correct rejections were similar across all conditions (see Figure 3.12). A one-way repeated-measures ANOVA revealed a main effect of face type, *F*(2, 32) = 8.73, *p* < .001, η_p^2 = .35. Tukey's HSD analyses were used to follow up the effect. Lower correct rejections were found for the two-face average (*M* = 66.38, *SD* = 18.16) and four-face average (*M* = 68.96, *SD* = 19.19) than the four exemplar images (*M* = 72.89, *SD* = 21.84), *ps* < .018. Correct rejection scores were comparable between the two-face and four-face averages, *p* = .664. Finally, a paired samples t-test revealed higher correct rejections for two-face averages when 100% of targets were present

(*M* = 76.17, *SD* = 13.91) compared to when 50% were present, *t*(17) = -2.69, *p* = .015, *d* = -.063. This was done solely for the two-face average conditions as this was the only condition where correct rejections were possible in the 100% trials. In the 100% of the four-face average and four exemplar image trials, all selection identities were matches, and so no correct rejections are possible. However, in the 100% of the two-face average conditions, there are only two matches in the four selection identities. The other two are therefore correct rejections.

Figure 3.12

Correct rejection rate across face types by percentage of matching targets present. In the 100% of four-face average and four image trials, all faces were matches, and so no correct rejections are reported. Error bars show standard error.



Discussion

The aim of this experiment was to determine if constituent identities of an average could be identified based on their resemblance to an encoded average. This was examined by introducing a memory demand, whereby observers were given the average (two- or four-face) or four exemplar images to encode before selecting the matching identities. Moreover, in contrast to previous studies, in which the formation of averages is assumed and the parallel encoding of exemplars cannot be ruled out (e.g., Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Neumann et al., 2018; Peng et al., 2022), participants were asked to encode averages directly.

Under these conditions, performance for the four-face average and four exemplar images were comparable, but both resulted in lower accuracy than the two-face average. This diverges from the previous experiments (Experiments 4 and 5), in which performance was lower for the four-face average in comparison to the two-face average or four exemplar images.

On one hand, this could provide support for ensemble coding, as performance was comparable for the four-face average and four exemplar images conditions, which suggests participants were encoding an average of the four exemplar images and this yielded similar results as the direct encoding of an average of the same identities. However, if participants are unable to see the resemblance between an encoded average and the constituent identities, then the identification of the average alongside the exemplar in Chapter 2 and in the literature (e.g., Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019) cannot be attributed to the average resembling the exemplar at a level sufficient to facilitate recognition. This therefore would suggest that participants were identifying the average because an average had been encoded.

An alternative argument is that the exemplar is encoded alongside the average (Li et al., 2016; Neumann et al., 2018). However, in the present study, resemblance between the test faces and the exemplars of the four exemplar images condition could not be easily detected with the memory component, suggesting that the individual identities were not encoded sufficiently. This could be the result of the difficulty of the task, as even under optimised conditions (e.g., unlimited time, simultaneous presentation of selection identities and the four exemplar images), Experiment 5 demonstrated that resemblance is difficult to detect between the four exemplar images and the matching identities, with only 74.6% accuracy. With a memory demand, both average conditions would have an advantage over the four exemplar images condition, as there was only a single face to remember in the average conditions, whereas the four exemplar images required the encoding of four faces. It is therefore possible that the lower accuracy for the four exemplar images trials in this experiment in comparison with the average conditions is due to the additional demands of this condition.

Experiment 7

Experiment 6 demonstrated that adding a memory demand to the identification task resulted in comparable accuracy between the four-face averages and four exemplar images conditions. This differed from Experiments 4 and 5 which found selectively lower accuracy for the four-face average condition. Whether this is due to participants forming an internal average of the four exemplar faces in Experiment 6, or whether this can be attributed to more difficulty in recalling four identities, is unclear. Previous research suggests that face processing is subject to a strict capacity limit, with one face being processed at a time (Bindemann et al., 2005). Therefore, the four exemplar images condition, which requires participants to encode four faces, might have been at a disadvantage compared to the averages, which contained multiple identities but only required the encoding of a single face.

This experiment aims to investigate this possibility by introducing a high- and low-load visual working memory task between encoding and identification. Working memory tasks between encoding and retrieval disrupt the retention of encoded items (e.g., Cowan et al., 2007). If the comparable performance of the four exemplar images and the four-face average conditions in Experiment 6 can be attributed to increased memory demands, then the high-load visual working memory task could be expected to disproportionately impact the four exemplar images trials. However, if the findings of Experiment 6 were due to the average representation not containing sufficient identity information to identify the constituent identities, and the exemplars of the four exemplar images condition were encoded into an average that is comparable with the four-face average, then the difference between these conditions should not be affected by the working memory load manipulation.

Methods

Participants

Thirty-four participants (6 male, 27 female, 1 other) were recruited in exchange for 4 credits for their course. Ages ranged from 18-33 years (*M* = 19.82, *SD* = 2.91). To partake in this experiment, participants must not have taken part in any of the previous experiments.

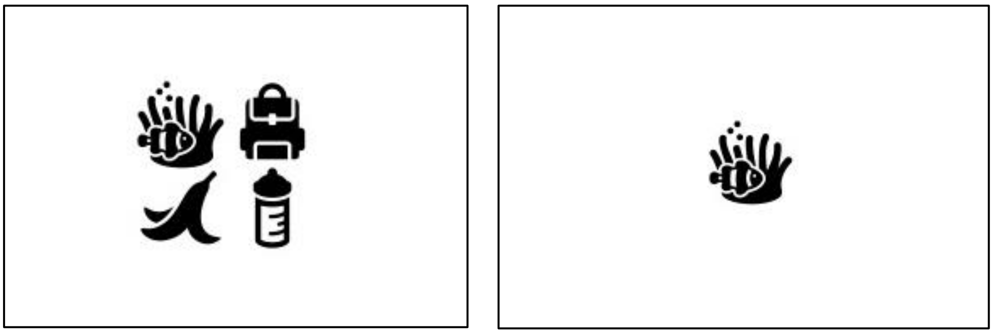
Methods, Design, and Procedure

The design followed the same structure as Experiment 6 where participants viewed an average of two or four faces or four exemplar images before being shown the probe display. Participants were shown either four images of different identities, an average of four identities, or an average of two identities for 4.5 seconds before a

1 second fixation cross. To manipulate memory demands, a visual working memory task was included in between face encoding and identity selection from the probe display. There were two versions of the working memory task – a high-load and a low-load. In the high-load condition, participants were shown four objects for 3 seconds followed by another fixation cross for 1 second. For the low-load condition, only two objects were shown in this timeframe. Participants were then presented with a single object until a response was registered and asked to press 'S' if the object was present in the previous display or 'L' if the object was not present. Objects were sourced from Microsoft Icons and were black and white graphics of various objects and animals (e.g., an acorn, a fish, an apple; see Figure 3.13). Overall, there were 48 object learning displays, of which 24 were four object displays and the remaining 24 were two object displays. Half of these trials were match trials, whilst the other half were mismatch trials in which a new object was displayed. This resulted in a total 168 objects. Object learning displays were shown in a grid of approximately 103 (w) x 108 (h) pixels, with a resolution of 54 ppi. Recognition objects were shown at approximately 60 (w) x 55 (h) pixels, with a resolution of 54 ppi.

Figure 3.13

An example of the object memory task. Left panel shows a learning display. Learning displays contained either two objects (low load) or four objects (high load). Right panel shows the test display.



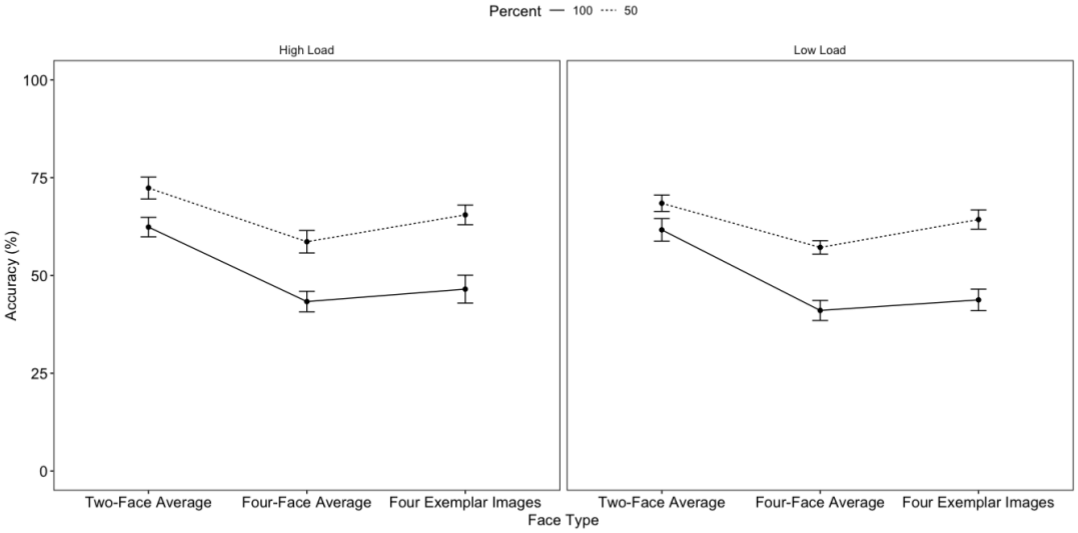
After another fixation cross of 1 second, participants were presented with the probe display and asked to select which of the faces they believed were used to create the learning display. Following this, participants completed the familiarity check for each identity. There were 48 trials with a break halfway through. The experiment lasted approximately 60 minutes.

Results

Overall accuracy on the working memory task was at 93.87%. A paired samples t-test showed that the memory load manipulation was successful, with higher accuracy on the low-load trials (*M* = 96.67%, *SD* = 8.68) than on the high-load trials (*M* = 91.07, *SD* = 9.39), *t*(34) = 5.46, *p* < .001, *d* = .92. The identity task data was cleaned by removing any trials that included a face a participant reported as familiar. This resulted in the removal of 623 trials (37.08% of all trials, *Mode* = 0%, *M* = 17.80%, *SD* = 15.83%). Overall accuracy for the identity task was then analysed and was at 57.27%. Highest accuracy was observed in the two-face average condition, followed by the four exemplar images condition, with lowest accuracy for four-face average trials (see Figure 3.14). Moreover, higher overall accuracy was seen for 50% trials compared to 100% trials.

Figure 3.14

Graph showing accuracy across each face type by percentage of matching targets present. Error bars show standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar image) x 2 (Target Presence: 50% vs 100%) x 2 (Memory Load: Low vs. High) repeated-measures ANOVA was then used to analyse the data. There was a main effect of Face Type, $F(2, 40) = 29.93$, $p < .001$, $\eta_p{}^2 = .60$, and Target Presence, $F(1, 20) = 42.30$, $p < .001$, $\eta_p{}^2 = .68$, but no main effect of Memory Load ($M_{Low} = 56.19$, $SD_{Low} = 16.60$; $M_{High} = 58.24$, $SD_{High} = 18.15$), $F(1, 20) = 0.08$, $p = .776$, $\eta_p{}^2 = .00$. The effects of Face Type and Target Presence were qualified by an interaction, $F(2, 40) = 6.95$, $p = .003$, $\eta_p{}^2 = .26$. Tukey's HSD revealed higher accuracy on two-face average trials ($M = 66.04$, $SD = 14.74$) than on four exemplar images ($M = 55.09$, $SD = 18.07$) or four-face average trials ($M = 49.94$, $SD = 15.31$), *p*s < .001. Four exemplar image trials did not result in reliably higher accuracy than four-face average trials, p = .056. In addition, 50% trials (M = 64.55, SD = 13.98) were found to produce higher accuracy rates than 100% trials (M = 50.17, SD = 17.51), p < .001. When only 50% of identities were present, accuracy was higher for two-face averages than four-face averages, p = .006 (see Table 3.7). For two-face averages, comparable accuracy was found with 50% of targets present as with 100%, p = .663. However, accuracy was higher on 50% than 100% trials for both four-face averages and four exemplar images, *ps* < .001. With 100% targets present, accuracy for two-face averages was higher than both four-face averages and four exemplar images, *p*s < .001. No other comparisons were significant, *ps* > .079. Finally, no interactions were found between Face Type and Memory Load, F(2, 40) = 0.30, p = .740, $\eta_p{}^2$ = .01, Target Presence and Memory Load, F(1, 20) = 0.08, p = .781, $\eta_p{}^2$ = .00, and between all three factors, F(2, 40) = 0.51, p = .602, $\eta_p{}^2$ = .03.

Table 3.7

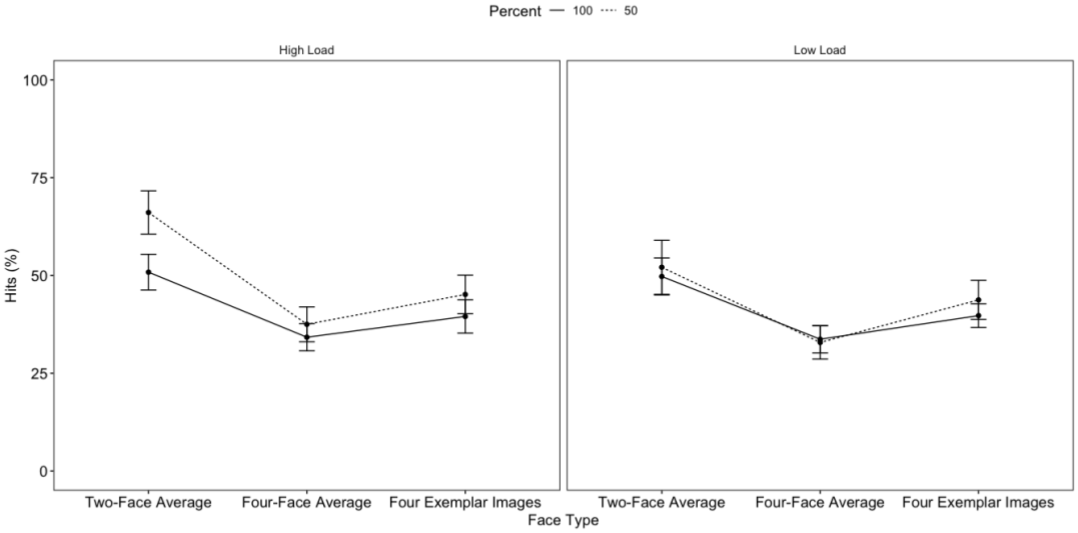
	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
			Low Memory Load			
Mean	68.45	61.65	57.17	41.05	64.29	43.75
SD	11.12	16.3	8.59	13.27	13.03	14.76
			High Memory Load			
Mean	72.34	62.36	58.62	43.32	65.49	46.5
SD	15.04	13.62	15.48	14.12	13.78	18.87

Hits

Data was then split into hits and correct rejections. For hits, a similar pattern to overall accuracy was found, with higher hit rates on two-face average trials followed by four exemplar image trials, and lowest accuracy on four-face average trials (see Figure 3.15).

Figure 3.15

Graph showing hit accuracy across each face type by percentage of matching targets present. Error bars show standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar images) x 2 (Target Presence: 50% vs. 100%) x 2 (Memory Load: Low vs. High) repeated-measures ANOVA was used to analyse hits. There was a main effect of Face Type, $F(2, 46) = 11.90, p < .001$, $\eta_p^2 = .34$. Tukey's HSD showed that two-face averages ($M = 54.55$, $SD = 30.03$) resulted in higher hits than both four-face averages ($M = 34.59$, $SD = 20.91$) and four exemplar images ($M = 42.02$, $SD = 23.50$), *p*s < .041, but no difference was observed between four-face averages and four exemplar images, $p = .220$. A main effect of Target Presence ($M_{50\%} = 46.37$, $SD_{50\%} = 29.65$; $M_{100\%} = 41.39$, $SD_{100\%} = 22.74$), $F(1, 23) = 1.65, p = .212$, $\eta_p^2 = .07$, or Memory Load ($M_{Low} = 42.15$, $SD_{Low} = 26.01$; $M_{High} = 45.51$, $SD_{High} = 26.76$), $F(1, 23) = 2.12$,

p = .159, η_p^2 = .08, was not found. Moreover, there were no interactions between Face Type and Target Number, F(2, 46) = .013, p = .875, η_p^2 = .01, Face Type and Memory Load, F(2, 46) = 0.47, p = .626, η_p^2 = .02, and Target Number and Memory Load, F(1, 23) = 3.26, p = .084, η_p^2 = .12. In addition, a three-way interaction was not observed, F(2, 46) = 0.56, p = .573, η_p^2 = .02 (see Table 3.8).

Table 3.8

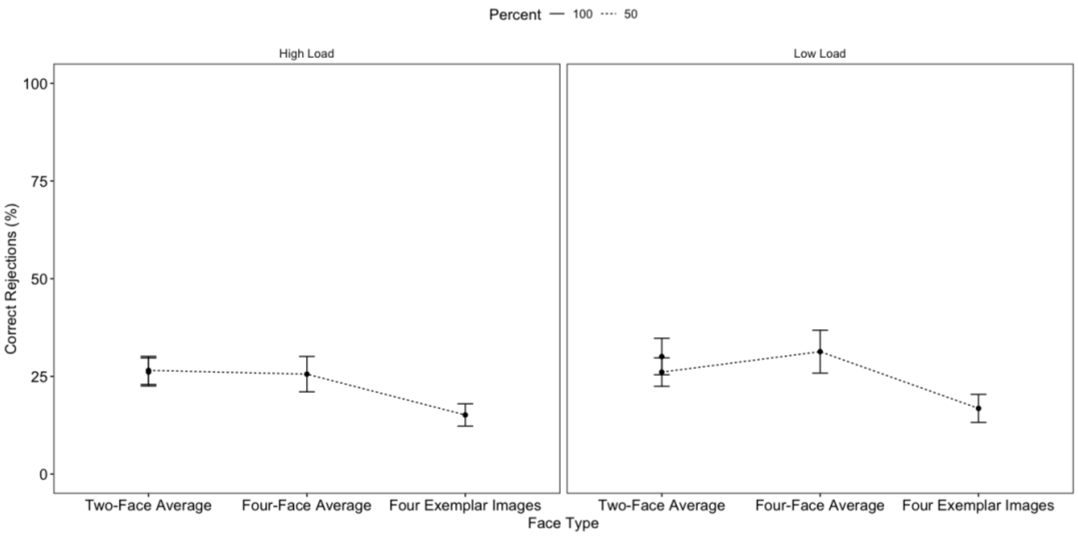
	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
			Low Memory Load			
Mean	52.08	49.74	32.87	33.69	43.75	39.73
SD	36.54	26.77	22.12	18.73	26.41	16.3
			High Memory Load			
Mean	66.09	50.83	37.5	34.21	45.14	39.51
SD	29.79	24.99	24.01	19.37	27.05	23.36

Correct Rejections

Correct rejections were found to be lowest for the four exemplar image trials, across both low and high memory load conditions. Both the two-face average and four-face average conditions showed comparable correct rejections, unlike the pattern seen for the overall accuracy and hit rates (see Figure 3.16).

Figure 3.16

Graph showing correct rejection rate across each face type by percentage of matching targets present. In the 100% of four-face average and four exemplar image trials, all faces were matches, and so no correct rejections are reported. Error bars show standard error.



A 3 (Face Type: two-face average, four-face average, four exemplar images) x 2 (Memory Load: Low vs. High) repeated-measures ANOVA was used to analyse the correct rejections. This revealed a main effect of Face Type, *F*(2, 46) = 3.37, *p* = .043, η_p^2 = .13, but Tukey's HSD found no comparisons between conditions to be significant ($M_{Two\text{-}Face}$ = 26.32, $SD_{Two\text{-}Face}$ = 19.16; $M_{Four\text{-}Face}$ = 28.35, $SD_{Four\text{-}Face}$ = 26.36; $M_{Exemplar}$ = 15.95, $SD_{Exemplar}$ = 17.14), *p*s > .065. A main effect of Memory Load (M_{Low} = 24.67, SD_{Low} = 23.10; M_{High} = 22.33, SD_{High} = 20.51), *F*(1, 23) = 0.04, *p* = .848, η_p^2 .00, and an interaction of Face Type and Memory Load were not found, *F*(2, 46) = 0.50, *p* = .612, η_p^2 = .02 (see Table 3.9).

Table 3.9

	Two-Face Average		Four-Face Average		Four Exemplar Images	
	50%	100%	50%	100%	50%	100%
			Low Memory Load			
Mean	26.09	30.08	31.33	-	16.82	-
SD	19.26	26.41	28.48	-	18.92	-
			High Memory Load			
Mean	26.53	26.11	25.57	-	15.14	-
SD	19.4	19.57	24.39	-	15.57	-

Discussion

This experiment aimed to investigate whether higher memory demands could explain the comparable performance for the four-face average and four exemplar images seen in Experiment 6. To test this, two manipulations of a working memory task were included in-between learning the set and identifying the constituent identities – a high memory load task and a low memory load task. It was expected that if performance on the four exemplar images condition could be attributed to remembering four faces as opposed to a single average face, then performance for this condition should be disproportionately affected by the load of the working memory task. Contrary to this expectation, no effect of memory load was found, with a comparable pattern of results found for both low load and high load trials.

Despite this, the results demonstrated a similar pattern to the effects seen in Experiment 6, where similar accuracy rates were observed for the four-face averages and the four exemplar images. Additionally, both four exemplar images and four-face averages were found to result in lower performance than two-face averages. This differs from Experiment 5, where comparable accuracy was observed between two-face averages and four exemplar images, with lower accuracy for four-face averages.

It is possible the working memory task was not suitable to interfere with the memory for faces. Some studies have shown that face memory is specifically affected by when working memory is loaded with other facial information (Cheung & Gauthier, 2010; Park et al., 2007; Wammes & Fernandes, 2016). Therefore, the use of objects in the working memory task may not have provided suitable interference to generate a memory load effect on the face recognition task. On the other hand, if a face working memory task has been used, then this might have interfered with the encoded representation. Ensemble coding has been found to be modifiable and updated by temporal associations (e.g., Haberman et al., 2009; Yang et al., 2024). Therefore, using a face working memory task may have inadvertently changed participants' encoded representation, making it harder to use resemblance to gain insight into the internal representation.

Nevertheless, memory load had no effect on the ability to spot the resemblance between either the average or the individual images and their constituent identities. This suggests that identifying the constituent identities for a four-face average is simply a hard task. When a memory demand is introduced, and participants have to learn four images, it is possible they are then forming an average representation, which causes an impairment in the ability to recognise resemblance to their constituent identities.

General Discussion

The purpose of this study was to investigate the extent to which people can spot the resemblance between an average and its constituent identities. Experiment 4 required participants to decide which of four faces were used to create a four-face average shown to them simultaneously. Performance on this task was poor at around 52% showing that it is difficult to spot the resemblance between a four-face average and its constituent identities. This provides *indirect* support for the ensemble coding of identity. If the average was being recognised because it resembled the encoded exemplar enough to be mistaken as a different image of that exemplar then observers should be able to detect the resemblance of an average to its constituent identities. However, these studies suggest that this is not the case, and the average does not demonstrate sufficient resemblance to the exemplars to be recognised off this alone.

Instead, the average must be selected because it resembles an internal representation – an ensemble.

Experiment 5 investigated this further by comparing the ability to identify the constituent identities of two-face averages and four exemplar images to that of four-face averages. Results indicated that spotting resemblance is more difficult for a four-face average, with better performance on both two-face average and four exemplar images trials. This implied that the two-face average sufficiently resembles the targets, and that the difficulty in detecting resemblance between the four-face average and its constituent exemplars was not due to the number of comparisons, as performance for the four-images matched that of the two-face average despite having the same comparisons as the four-face average.

Memory is an important element of ensemble coding, however, and therefore Experiment 6 added a memory demand to study the effect this would have on the identification of the constituent identities. With this memory demand, performance was poor for both the four-face average and the four exemplar images but was higher for the two-face average. This could be attributed to the higher memory demand for the four exemplar images condition compared to the single face learned in the average conditions. Moreover, performance for the four-face average was still lower than the two-face average, suggesting that encoding the average did not change the lack of resemblance to the constituent identities.

To test if the memory effect seen in Experiment 6 could be attributed to having to learn four images as opposed to a single average face, Experiment 7 introduced a visual working memory task. If participants had difficulty learning four identities, a high load task of visual working memory should selectively interfere with accuracy for the four exemplar images. Alternatively, if participants were forming an ensemble and unable to identify the constituent identities because of this, the working memory task should not have affected accuracy for the four exemplar images any more than that of the four-face average. Results provided partial support for the ensemble explanation, as there was no observed difference in performance between the four-face average and the four exemplar images. However, working memory load had no effect on the results. This corresponds to previous findings that ensemble coding is robust to working memory load interference (Peng, Kuang, et al., 2019). For instance, having to determine if a target orientation is present among a series of distractors in between encoding the faces and the recognition test results in comparable ensemble effects regardless of whether the distractors are congruent or incongruent to the target (Peng, Kuang, et al., 2019).

Overall, the results from this study show that the resemblance between a four-face average and its constituent identities is small and difficult for observers to detect *directly* in a visual comparison. This supports previous studies on the ensemble coding of identity, which often accept equal recognition of the exemplars and averages as evidence of ensemble coding (e.g., Neumann et al., 2013; Neumann et al., 2018; Peng, Zhang, et al., 2019; Peng et al., 2021). The results from Chapter 2 suggested that averages could be recognised based off resemblance to the encoded identities rather than the formation of an internal average. However, previous studies typically use four or more identities in their encoding sets (e.g., Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al., 2022; Sama et al., 2019). This chapter shows that averages made from four or more identities are unlikely to be recognised based on resemblance to the encoded exemplars. Therefore, the comparable recognition of exemplars and averages (e.g., Neumann et al., 2013; Neumann et al., 2018; Peng, Zhang, et al., 2019; Peng et al., 2021) must be because the internal representation *is* an average.

The formation of an internal average may be related to how it is thought that familiar faces are recognised. Whilst unfamiliar face matching is a difficult task (e.g., Bruce et al., 2001; Megreya & Burton, 2006b; Ritchie et al., 2022; White et al., 2014), familiar face matching is much more accurate (e.g., Bruce et al., 2001; Clutterbuck & Johnston, 2005). There is strong evidence that this is because multiple exposures to the face enables a stable representation to be formed (Corpuz & Oriet, 2022; Ritchie & Burton, 2017; Sandford & Ritchie, 2021). Current theorising suggests that this representation takes the form of an average, whereby each encounter with the face refines the representation and extracts consistent identity-specific information whilst removing variable image characteristics (Jenkins & Burton, 2011). Whilst this is assumed to be identity-specific, humans frequently make errors in identity matching. For instance, different identities are mistaken as the same person approximately 25% of the time (Burton et al., 2010; Ritchie et al., 2022). This can be influenced by participants' expectations. For instance, in paradigms where observers are exposed to multiple images of an identity before being tested on a novel image, recognition accuracy for the novel image is improved if participants *believe* the images to show a single person (Menon et al., 2015; Menon et al., 2018). Likewise, if participants believe the images to depict two different identities, accuracy is reduced (Menon et al., 2015; Menon et al., 2018). Additionally, research on associative priming demonstrates that recognition of a target can be facilitated if preceded by an associated face (Stevenage et al., 2014; Vladeanu et al., 2006). These associations can be formed based on semantic information, such as occupation or location, but also based on co-occurrence

(Vladeanu et al., 2006). It is therefore possible that the co-occurrence of faces in an encoding set of an ensemble paradigm results in the use of the same mechanisms as those used to create the average face representations, resulting in the formation of an ensemble average.

Whilst this experiment provides the most persuasive evidence that an average of identity is being encoded, this leaves the issue of what causes the average to be chosen over the exemplar. In some studies, there is evidence that the average is selected at a more frequent rate than the exemplar (Matthews et al., 2018; Peng, Zhang, et al., 2019; Rhodes et al., 2015), whilst in others, the average and exemplar are selected at comparable frequencies (Bagaïni & Hole, 2017; Peng et al., 2021; Peng et al. 2022; Sama et al., 2019). The results in Chapter 2 showed comparable selection of the exemplar and the average (Experiments 1-2). This remained the case when participants were asked to decide *between* these faces (Experiment 3). Despite this, participants struggle to see the similarity between a four-face average and an exemplar, suggesting the representations are independent. The question remains as to when does the average have an advantage over the exemplar if not based off similarity to the encoding identities?

One possibility is that humans do not have an objective perception of similarity. For example, humans have been shown to be inconsistent in their own judgements of similarity, making different judgements to the same items on different days (Bindemann, Avetisyan, et al., 2012). Therefore, observers' ability to spot resemblance may not be an adequate measure of the importance of similarity in recognising the average. A more consistent and accurate similarity measure is therefore required for this. One potential source of this is facial recognition algorithms. Although there was original doubt on these, recent work has shown that these can outperform human observers (Phillips et al., 2018). Therefore, to explore what representation is given more weighting, Chapter 4 will use similarity ratings given by a facial recognition algorithm and compare this to which face - the average or exemplar - participants chose in Chapter 2 Experiment 3.

CHAPTER 4:

Algorithm Similarity Ratings of an Exemplar to an Average

4.1. Introduction

The experiments of Chapter 3 consistently demonstrate that recognition of the constituent identities of a face average, which was created from four separate faces, is a difficult task. When the average was shown simultaneously with a set of selection faces, which comprised of a mixture of the average's constituent identities and some foil identities, recognition performance was only around 51.6% (Experiment 4). Accuracy was higher for averages created from two faces compared to the four-face average, which demonstrates that the detection of resemblance with a constituent selection face becomes more challenging as more identities are added to an average (Experiment 5). In addition, detecting the resemblance for four individual exemplar images to the set of four selection identities was comparable to performance with a two-face average. This shows that the disadvantage for the four-face average was not a consequence of the number of identities that were required to be processed, but the amount of identity information that was retained in an average (Experiment 5). The advantage for the two-face average remained when a memory demand was introduced, but this eliminated the advantage for the four exemplar images (Experiment 6). This might reflect the higher memory demands involved in processing four faces as opposed to a single average face, yet a high working memory load manipulation in-between learning and test did not impact memory for the four exemplar images any more than a low memory load (Experiment 7). Taken together, these experiments provide evidence in favour of ensemble coding of identity, as averages made from four or more faces are unlikely to be recognised based off their direct resemblance to the encoding images. These averages must therefore be recognised based on their resemblance to an *encoded* representation that combines facial information of its constituent identities.

While these findings indicate that similarity between individual face identities and an average is difficult to detect, research has consistently demonstrated that similarity is important for face recognition. For example, observers can readily detect resemblance between faces – known as between-person similarity (e.g., Gawronski & Quinn, 2013; Megreya & Burton, 2008; Megreya et al., 2013), even across different likenesses. Likeness refers to the degree a particular image of a face demonstrates the recognisable characteristics of that person. For example, in a study by Stephan and Arthur (2006), an expert and a novice were each asked to create a facial approximation from casts of the same skull using an image of that person. This resulted in two different approximations of the same person. Yet, resemblance ratings for these two approximations to the photo were only one point apart on a 0-5 Likert scale, despite different appearances (Stephan & Arthur, 2006). Resemblance is also used to aid recognition of unfamiliar identities, where an unfamiliar face that resembles a known face can be better remembered than one that does not (Tomita et al., 2014). This resemblance can also be strong enough to cause individuals to make mistaken identifications (Megreya & Burton, 2008; Megreya et al., 2013). The body of research on face matching and eyewitness testimony highlights the prevalence of such misidentifications (e.g., Jones et al., 2020; Megreya et al., 2013; Rose & Beck, 2016; Thompson & Johnson, 2008; White et al., 2014), underlining the issues resemblance can cause for unfamiliar faces.

In addition to between-person similarity, faces also demonstrate within-person variability. For instance, the same face can exhibit substantial variation in appearance under different lighting conditions or viewing angles (Braje, 2003; Lee et al., 2006). Moreover, faces also change over time, for instance, through ageing or changes in adiposity (Longmore et al., 2017). This variability is idiosyncratic (e.g., Burton, 2013; Jenkins et al., 2011), with different identities showing different ranges of variability. Additionally, the amount a specific instance will resemble a single identity can be both strong and weak. What can be considered a strong resemblance depends on observers' previous experience with that identity (Ritchie et al., 2018). For instance, if a person is only ever encountered in one context (e.g., at work), then the representation of that person will be based on that specific context (e.g., clean-shaven, make-up). An instance of that individual outside of that context (i.e., flushed and sweaty at the gym) will be considered poor likeness to the internal cognitive representation of the identity. However, for other people, who encounter that individual within this context, this may be considered a good likeness. This poses a significant challenge in understanding resemblance, as is within-person variability idiosyncratic (i.e., Burton, 2013; Jenkins et al., 2011) *and* based on an observer's specific perception of resemblance.

There is also evidence to suggest that individuals are not consistent with their judgements of similarity (Alenezi et al., 2015; Bindemann & Sandford, 2011; Bindemann, Avetisyan, et al., 2012; Russ et al., 2018). On different days, the same observers will make different judgements to the same pairs of faces (Bindemann, Avetisyan, et al., 2012), suggesting observers do not use similarity in an objective or consistent manner. On the other hand, observers' judgements of facial similarity correlate with their identification accuracy (Fysh & Bindemann, 2023). This creates a paradox in which similarity drives identifications but can be unreliable and used in an inconsistent manner.

Whilst the effects of similarity on facial identifications have been researched intensively (e.g., Bicego & Grosso, 2019; Fitzgerald et al., 2013; Honig et al., 2022; Lucas et al., 2021), how observers process this similarity has not. One view is that similarity uses holistic processing, in which faces are processed at a glance and as an integrated *Gestalt*, whereas featural information is used for judging dissimilarity (Bicego & Grosso, 2019). Another possibility is that more frequent use of featural information, and a reduction in holistic processing, may account for perceived similarity between faces. For instance, the other-race effect (ORE) describes a phenomenon that occurs whereby observers report higher similarity between faces of a different ethnicity than between faces of their own ethnicity (e.g., Chang et al., 2015). One explanation for this effect is that holistic processing is reduced for outgroup faces (e.g., DeGutis et al., 2013; Tanaka et al., 2004), giving more weight to featural information. It is therefore possible that this higher emphasis on featural processing increases perceived similarity for outgroup faces. Moreover, the literature on kinship detection also demonstrates that observers use features in the upper half of the face, and particularly the eye regions, to make kinship judgements (e.g., Arantes & Berg, 2012; Dal Martello & Maloney, 2006). Furthermore, two face templates that share the same features are rated as more similar than faces that share the same template but contain different features (Esins et al., 2011). Featural information is therefore more influential in the perception of similarity as changes to configural (holistic) properties of a face (i.e., the template) do not affect similarity judgements if the features remain the same. Taken together, these studies favour a featural account of similarity, with reduced holistic processing.

One model of similarity defines it as the sum of the similarities and differences of the features of two items (Tversky, 1977). To do this, the similarity of each feature for two items needs to be calculated. For faces, this involves calculating a similarity score for each facial feature (e.g., eyes, nose). Once these have been collected, overall similarity can be computed from the average of these similarity ratings. There is evidence that the average similarity across facial features correlates with match accuracy, however, this seems to be less informative of mismatch accuracy (Fysh & Bindemann, 2023). Moreover, if similarity was objectively calculated from features, then the perception of similarity should be consistent both within- and across-observers, which contrasts with the individual differences within and between observers that have been reported in the literature (e.g., Bindemann, Avetisyan, et al., 2012; Ritchie et al., 2018). Therefore, to fully understand the effects of similarity both within- and across-identities and observers, a subjective measure from humans is not sufficient. A more consistent measure is required.

One such measure comes from facial recognition algorithms. These work by identifying a face within an image by looking for patterns that resemble a face. Once a face has been detected, it maps out distinguishing key landmarks (e.g., shape of the eyes, nose, mouth, etc.). This is then converted into a string of numerical values known as a faceprint. The faceprint for an image can be compared against another faceprint to determine if the identity is a match. Algorithms developed in 2015 have been shown to perform at comparable accuracy to human observers in some face matching tasks (Phillips et al., 2018). More recent algorithms perform at a level equivalent to and higher than the best human observers (e.g., Carragher & Hancock, 2023; Phillips et al., 2018).

This chapter takes advantage of the proficiency of facial recognition algorithms, by applying this to gain a consistent and sensitive measure of the similarity of average faces to their constituent exemplars. The question of main interest is whether this similarity correlates with the human data from Chapter 2, where observers had to select between the average or the exemplar (Experiment 3), and Chapter 3, where resemblance was investigated between a four-face average and its constituent identities (Experiment 4). If such a relationship is found then this provides evidence that observers are using similarity in a consistent manner, akin to that of a face recognition algorithm. In contrast, if such a relationship is not found, then this would suggest that the recognition of an average involves more than the processing of similarity to its exemplars. This would suggest that any recognition of an average in an ensemble coding paradigm must arises from the formation of internal average during the encoding of its constituent exemplar faces – otherwise the average should not be selected due to its lack of similarity to these faces.

To investigate this question, this chapter first examines algorithm performance for two well-established tests of face matching ability (i.e., Kent Face Matching Test, Glasgow Face Matching Test) to demonstrate its capability in detecting similarity and making identification decisions. To determine the level of similarity that a four-face average has to its corresponding constituent images, the algorithm will then be used to generate similarity ratings of these constituent identities from Chapters 2 and 3 to the corresponding four-face average. The similarity ratings from the algorithm for these face pairings will then be correlated with human performance from the experiments in Chapter 3. Similarity ratings will also be collected for two-face averages, and the ratings correlated against human performance from Chapter 2. If the algorithm has difficulty in detecting the similarity between the constituent faces to a two- or four-face average, then it is unlikely that sufficient similarity exists between an average and its constituent identities to allow for identification. Therefore, similarity scores would not be expected to correlate with human performance. However, if the algorithm detects high levels of similarity for a two- or four-face average and its constituent faces, then it is possible that this provides information that human observers are able to use in their identification decisions.

4.2. Efficiency of Facial Recognition Algorithms

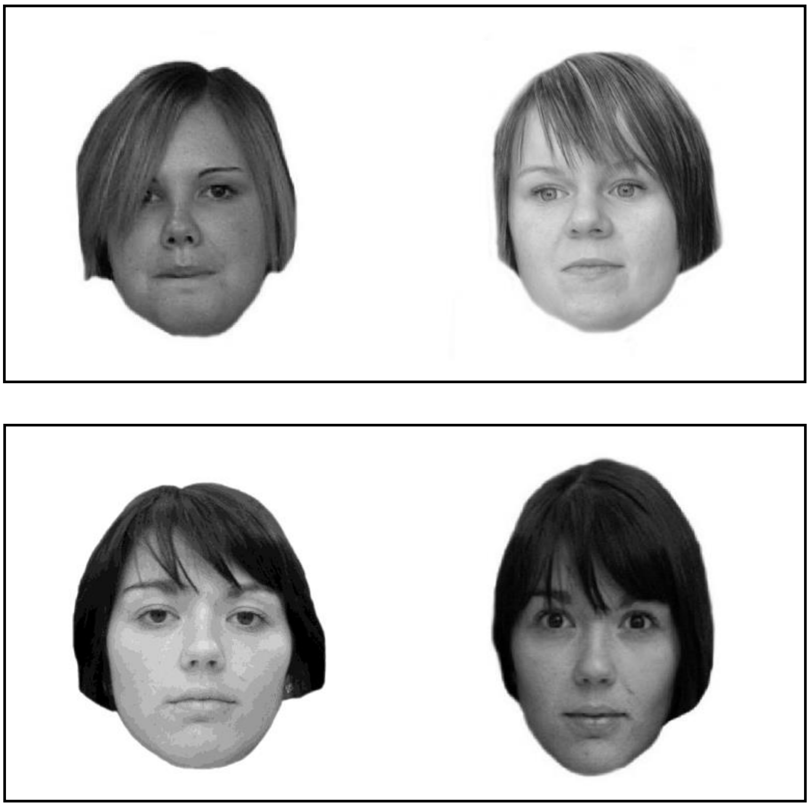
Facial recognition algorithms have been shown to outperform humans on challenging tasks (e.g., O'Toole et al., 2007; O'Toole et al., 2012; Phillips et al., 2018; Phillips & O'Toole, 2014). Because of this, algorithms are now employed routinely in applied settings. For instance, automatic face recognition systems are now used at border control, with over 290 e-gates in place in the UK (Neal, 2023). Facial recognition algorithms are also employed for identity verification for drivers' licences and for voter registration (see Huang et al., 2005). However, these systems are not infallible. For example, algorithm performance suffers when image quality is poor or if faces are masked (e.g., Jeevan et al., 2022; Zhou et al., 2018). Despite this, even under challenging conditions such as with cross-race comparisons and highly challenging items, algorithms perform better than humans (Jeckeln et al., 2023; Phillips et al., 2018).

A highly proficient face recognition algorithm that is publicly available is Amazon Rekognition. This is a deep-learning visual analysis algorithm which uses convolutional neural networks (CNNs). One feature of this application is facial image comparison, which enables the identity matching of two facial images. Once each face has been extracted and the features mapped into a feature space, Amazon Rekognition uses distance metrics (e.g., Euclidean distance, cosine similarity) to provide a similarity score, with a lower distance equating to a higher similarity score. If this score passes a similarity threshold (86%), then the two faces are considered a match. To demonstrate the accuracy of Amazon Rekognition, two well-established tests of face matching ability were run through the algorithm.

The first one of these tests is the Glasgow Face Matching Test – Short (GFMT-S; Burton et al., 2010). This test involves the comparison of two images to determine if they are an identity match or mismatch. The short version of this test contains 40 pairs of faces, with 20 male and 20 female face pairs. All faces are presented in a frontal position with a neutral expression (see Figure 4.1). Half of the trials are match trials, and the other half are mismatch trials. For each identity, two photographs were taken on two different cameras. Therefore, on match trials, the pair comprised of two different images of the same identity. Non-match trials were selected based on similarity ratings between identities from a pilot sorting trials (Bruce et al., 1999). Each image was grayscaled and resized to 350 pixels (w) with a resolution of 72 ppi (see Figure 4.1). Human performance on this test is approximately 89.9%, with 92% accuracy on match trials and 88% accuracy on mismatch trials (Burton et al., 2010). By comparison, across both match and mismatch trials, Amazon Rekognition produced 100% accuracy on this test. In addition, the mean similarity score provided by the algorithm was 99.99% for matches (range = 99.97 to 99.99, *SD* = 0.00) and 6.27% for mismatches (range = .002 to 26.12, *SD* = 8.47). Thus, whereas the GFMT-S provides optimal conditions for face matching, human observers still make substantial errors on this task. In contrast, Amazon Rekognition can perform this task with perfect accuracy.

Figure 4.1

An example of a mismatch face pairing (top panel) and a match pairing (bottom panel) from the GFMT-S.

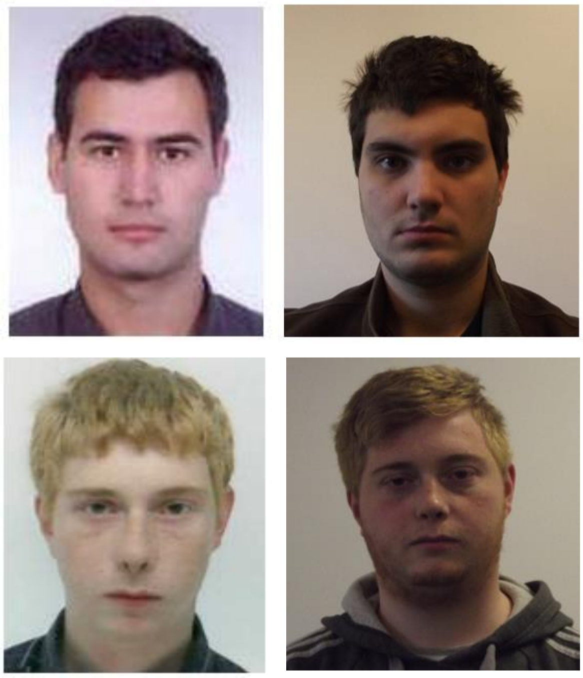


The second test that was employed to demonstrate the face matching ability of the Amazon Rekognition algorithm was the Kent Face Matching Test (KMFT; Fysh & Bindemann, 2018). This is a more difficult test than the GFMT-S and requires participants to compare a photo to a student ID card (see Figure 4.2). This test comprises of 40 pairs of faces (20 male, 20 female). Half of these are match trials, and the other half are mismatches. For each identity, two images were collected. One image was taken under controlled conditions with a frontal position and neutral expression. These were displayed at 283 x 332 pixels with a resolution of

72 ppi. The second image was taken from the subjects' student ID photograph. These were also frontal facing with a neutral expression but were taken under uncontrolled conditions. Student ID photographs were shown at 142 x 192 pixels with a resolution of 72 ppi. Mismatch pairs were selected based on their visual similarity using hair colour, eyebrow shape, and face shape. In the normative dataset for this test, human accuracy is at 66% for both matches and mismatches (Fysh & Bindemann, 2018). In contrast, Amazon Rekogition achieved 100% accuracy. Moreover, the mean similarity score provided by the algorithm was 99.72% for matches (range = 98.31 to 99.99, *SD* = 0.49) and 6.73% for mismatches (range = 0.13 to 37.59, *SD* = 10.35).

Figure 4.2

An example of a mismatch pairing (top panel) and a match pairing (bottom panel) from the KFMT.



Therefore, across both face tests, Amazon Rekognition produced perfect accuracy over both match and mismatch trials, substantially outperforming the mean accuracy of comparison groups of human observers. Indeed, while a small set of human observers can achieve 100% on the GFMT-S (e.g., Burton et al., 2010), no individuals scored above 90% in the normative sample for the KFMT (Fysh & Bindemann, 2018). Moreover, not only could the algorithm detect identity across different images taken on the same day (GFMT-S), but also with uncontrolled images that were recorded several months apart (KFMT). This suggests that the algorithm's ability to detect identity is sensitive to identity-specific information, robust to variability in images, and exceeds the ability of humans.

4.3. Algorithm Similarity Ratings of Averages

In Chapter 3, human observers had difficulty in matching identities to an average made up of four faces. In this paradigm, observers were shown the average alongside four selection faces and asked to determine which of these faces, if any, were used to create the average. Overall accuracy across a series of four experiments was approximately 56.9%, which shows that this is a challenging task. Even under the best conditions, for instance, with two images taken just minutes apart, face matching is a difficult task (e.g., Megreya et al., 2013). It is therefore not surprising that observers have difficulty matching a face to an average where identity not only has to be detected across a change in image, but is also intermixed with other identities.

Indeed, at this point it is unclear how much resemblance a four-face average carries to its constituent identities. A simple assumption is that an average of four faces, which were combined with a consistent method, also resembles each identity at a ratio of 1 to 4. Put differently, as each identity contributes 25% of the information contained in an average, the similarity of the average to each identity should be 25/100 on the similarity measure provided by Amazon Rekognition. However, the relationship of similarity and identity does not appear to follow such a simple order. For instance, humans are able to detect resemblance of an unfamiliar face to a familiar face despite knowing they are not the same person (e.g., Tomita et al., 2014). Moreover, averages constructed from two faces are often accepted as a match to one of their constituent identities and go undetected as combination of two identities (e.g., Heyer et al., 2019; Robertson et al., 2017), suggesting resemblance must be more than the sum of its parts. Observers therefore have to adopt a decision threshold whereby faces that reach *some* threshold of similarity are accepted as a match, and those that do not are classified as a mismatch. Depending on context, this threshold may be liberal, allowing the acceptance of quite dissimilar images, or conservative, where only images that are highly similar are accepted as a match (e.g., Baker et al., 2024; Stabile et al., 2024). However, even if it does not reach this threshold, this does not necessitate that similarity cannot be seen. It is unclear how much resemblance averages made from four identities have to each encoding face, yet the results from Chapter 3 show that humans struggle to detect this resemblance.

As Amazon Rekognition is more accurate than humans on tests of face matching ability (e.g., GFMT, KFMT), if the resemblance of a four-face average to its constituent identities is detectable, then this will be reflected in the similarity scores generated by this algorithm. To investigate this, the four-face averages and exemplars from Chapter 3 were compared against the corresponding encoding images with Amazon Rekognition. Each encoding image was also compared to a non-matching exemplar to compare the similarity of an average and its encoding image to that of a non-matching face.

On average, encoding images were rated as 99.46% similar to the corresponding exemplar faces (range = 29.93 to 99.99, SD = 4.77) with 99.17% of these being classed as a match (see Figure 4.3). By comparison, similarity for the non-matching exemplar was at floor, with a 1.54% average similarity rating (Min = 0.03, Max = 35.02, SD = 3.59), with no faces being classified as a match. This demonstrates the accuracy of the algorithm, as non-matching exemplars were not accepted as a match to the encoding images. In contrast, average similarity for the encoding images to the four-face averages was low at 36.29% (Min = 0.64, Max = 96.93, SD = 26.10), and only 3.33% of these stimuli were classified as a match.

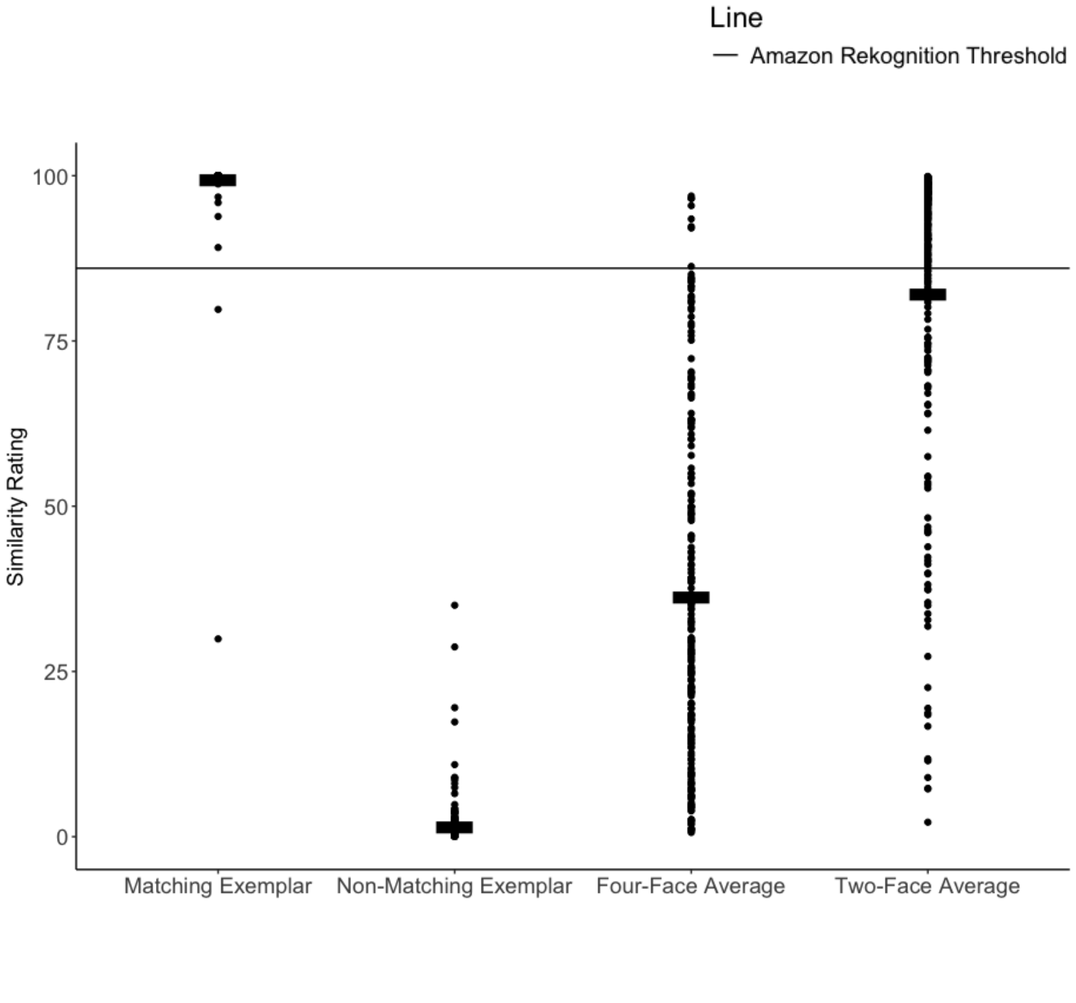
To analyse these data, a paired t-tests was conducted to compare the encoding images from the experiments in Chapter 3 with non-matching exemplars, matching exemplars and four-face averages. This showed that similarity was substantially lower for the encoding images and the non-matching exemplars compared with matching exemplars, $t(239) = 256.98$, $p < .001$, $d = 16.59$, and the four-face averages, $t(239) = 21.33$, $p < .001$, $d = 1.38$. In addition, four-face averages also bore lower similarity to the encoding images than their constituent exemplar faces, $t(239) = 37.51$, $p < .001$, $d = 2.42$. Taken together, these data reveal two key insights. First, Amazon Rekognition can determine reliably whether an exemplar has occurred in the encoding set. Indeed, Figure 4.3 illustrates the spread of similarity scores for individual faces and shows that all bar two exemplars were identified against the encoding images (i.e. above a similarity decision threshold of 86% that is applied by the algorithm), whereas similarity scores for non-matching exemplars were consistently low. In contrast, mean similarity ratings for four-averages were much lower than for matching exemplars and only a small number of these faces were matched against the encoding faces above threshold: whereas over 99% of exemplars were recognized as a match to the encoding images, only 3% of averages crossed this threshold.

Whilst four-face averages may not retain enough information from each constituent identity to be detected by humans or algorithms, averages made from just two faces should contain higher similarity, as there are less identities contributing to the average. In Chapter 2, observers were as likely to identify a two-face average as one of two targets as they were a different image of one of the targets. Moreover, in Chapter 3, whilst observers had difficulty identifying the constituent identities of a four-face average, accuracy was higher for two-face averages. From studies on face averaging, it is also apparent that participants can see the similarity between an average of two faces and its constituent identities (e.g., Kramer et al., 2019; Nightingale et al., 2021; Robertson et al., 2017). This similarity is strong enough that the average is often mistaken as an ambient image of one of the individuals. This can be seen in border security settings, where averages can be used as a means of fraudulent ID (e.g., Kramer et al., 2019; Robertson et al., 2017).

To examine the similarity of two-face averages to the constituent identities, Amazon Rekognition was again used to generate similarity scores between encoding images and two-face averages. Two-face averages were given an average similarity score of 82.04% to the corresponding encoding images (Min = 2.19, Max = 99.88, SD = 23.09), with 65.83% of these classified as matches (see Figure 4.3). A repeated-measures t-test showed that this was higher than the similarity of each encoding image to a non-matching exemplar, $t(239) = 55.19$, $p < .001$, $d = 3.56$, and higher than the similarity of a four-face average to the corresponding encoding image, $t(239) = 28.89$, $p < .001$, $d = 1.86$. However, this was lower than the similarity of the encoding images to the exemplar face, *t*(239) = 12.18, *p* < .001, *d* = 0.79, demonstrating the reduction in similarity when identities are averaged together.

Figure 4.3

Similarity scores generated by Amazon Rekognition for the encoding images to the matching exemplar, a non-matching exemplar, a two-face average, and the four-face average. Dots represent each similarity rating generated by the algorithm. The thick bars represent the mean for each image type.



These data demonstrate that two-face averages contain a relatively high level of similarity to the original encoding identities, and which is at a level just below the match threshold applied by the algorithm. Whilst the similarity threshold of an algorithm can be adjusted depending on its intended use, with more conservative thresholds for applications such as detecting identity fraud and more liberal thresholds in cases such as locating missing persons (e.g., Kanika et al., 2021; Shelke et al., 2021), Amazon Rekognition adopts a fairly conservative threshold, requiring a minimum of 86% similarity. Despite this, two-face averages contain enough similarity on average to almost reach this threshold, and more than a third of the two-face averages from this experiment crossed this identification threshold.

Overall, these results demonstrate that, on average, two-face averages retain high levels of similarity to their constituent identities. It is therefore possible that this similarity can explain the comparable selection of the exemplars and averages throughout Chapter 2. However, many other studies find comparable selection of the exemplar and average even when the average is comprised of four or more identities (Bagaïni & Hole, 2017; Neumann et al., 2013; Neumann et al., 2018; Sama et al., 2019). From the results shown here, four-face averages have less than half the similarity to the constituent identities than two-face averages. It is therefore unlikely that similarity to an exemplar can explain why a four-face average is selected as frequently as an exemplar (e.g., Bagaïni & Hole, 2017; Neumann et al., 2013; Neumann et al., 2018; Sama et al., 2019). This provides evidence that the average is selected because it matches an internal representation that is an average.

4.4. Do Algorithm Similarity Scores Correlate with Human Identification?

Despite the perfect performance of Amazon Rekognition on tests of face matching, the algorithm showed great difficulty in detecting the resemblance between the encoding images and the four-face average. In addition, the algorithm very rarely (8 out of 240 identities, equivalent to 3%) accepted four-face averages as a match to the original encoding identities. This demonstrates that detecting the presence of 25% of an identity in an average is difficult, as each identity in a four-face average contributes to 25% of the average. This converges with previous findings, which show that 20% of an identity can go undetected within an 80:20 average (Hsu & Lee, 2016; Jenkins & Burton, 2011). It is also consistent with the results from Experiment 4 in Chapter 3, which show that observers have difficulty in identifying the constituent identities for a four-face average. However, two-face averages contained higher similarity to each of its constituent identities and were accepted as a match more frequently (158 out of 240 identities/66%) than the four-face average.

From the results of Chapters 2 and 3, it is evident that humans must also be able to detect this resemblance. In Chapter 2 Experiment 3, human observers encoded two identities before selecting which of two faces most resembled one of the encoded identities. In this study, observers were more likely to select both the exemplar and the average faces when these were paired against new identities. When these were paired against each other, observers were as likely to select either the exemplar or the average. However, the question arises of whether human observers use similarity in a similar manner to the algorithm, considering that the algorithm provides higher similarity ratings for matching exemplars than two-face averages.

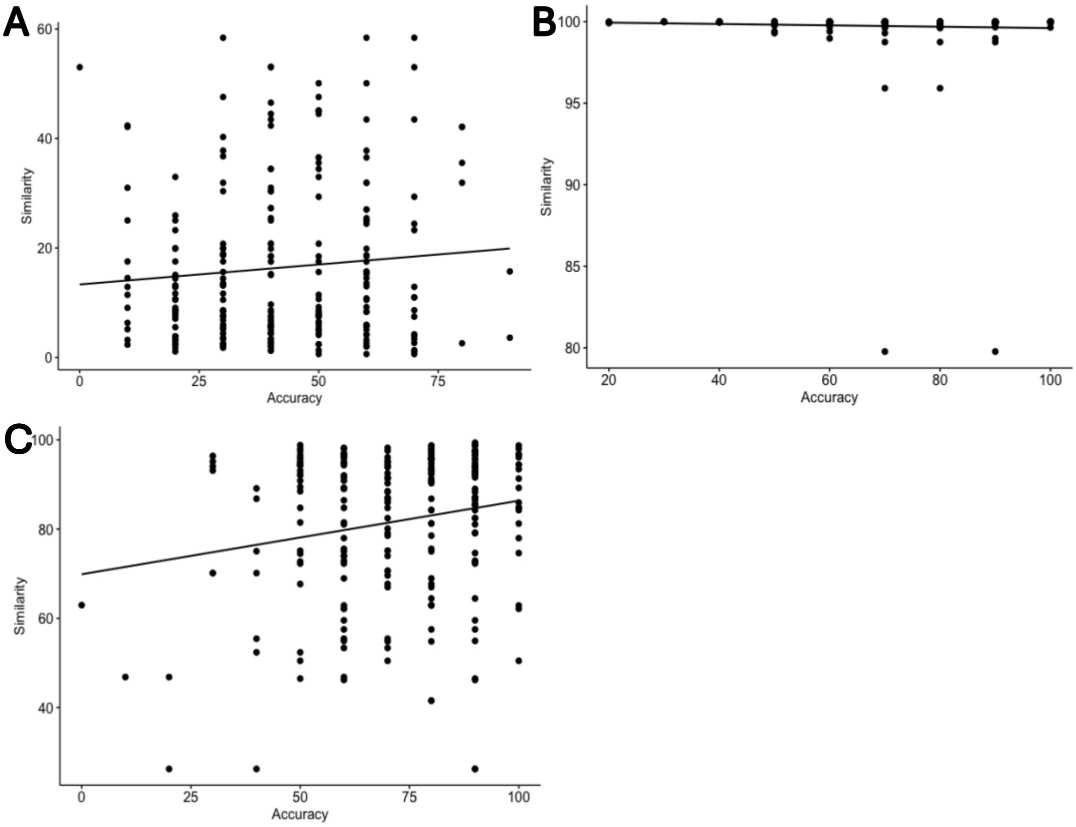
Whilst it is evident that humans have difficulty determining which face, the exemplar or average, is most similar to the encoded targets, this does not mean they cannot see the similarity. It is likely humans have a similarity threshold, whereby any instance that reaches this threshold of similarity to a target is considered a match. However, it is not clear what this threshold is. It is possible that humans have a lower threshold than Amazon Rekognition (86%). This may be reached by two-face averages, causing identification of both the exemplar and average in ensemble paradigms (e.g., e.g., Neumann et al., 2013; Neumann et al., 2018; Peng, Zhang, et al., 2019; Peng et al., 2021). However, this argument relies on the assumption that humans are using similarity in a similar way to an algorithm. Therefore, to investigate this, this experiment correlated human decisions between the exemplar and a two-face average with the similarity scores generated by the algorithm. If humans are able to see the similarity, and use this in their identification decisions, then human decisions should correlate with the similarity scores. On the other hand, if participants are not using similarity in the same way, then no correlation should be found.

To examine this, the similarity scores from Amazon Rekognition were correlated on a by-item basis against the human data from Experiment 3 of Chapter 2. A difference score was calculated for each item, which involved subtracting the similarity of the two-face average to the encoding image from the similarity of the exemplar to the encoding image. This was done to control for differences in similarity per trial (i.e., some trial had higher baseline similarity for both the average and exemplar which would result in a low difference score and vice versa), but also to test whether the differences in similarity could account for observer selection of each face. For example, higher difference scores could lead to more frequent selection of the exemplar, whereas lower difference scores could result in more mixed responses. This approach also addresses some of the issues with the limited range of similarity scores for the exemplar (difference range of 57.77 vs. exemplar range of 35.04).

When participants were asked to choose between the average and the exemplar, the difference in the similarity ratings of these faces to the encoding images did not correlate with the selection of the exemplar, *r*(238) = .141, *p* = .161 (see Figure 4.4A). That is, the difference in similarity between the matching exemplar and a two-face average in comparison with the encoding images, as detected by the Amazon Rekognition algorithm, did not relate to the likelihood that the exemplar was chosen over the average (or vice versa). Similarly, when the exemplar was paired against a new identity, similarity of the exemplar image to the encoding images was not found to correlate with identification of the exemplar, *r*(238) = -.04, *p* = .526 (see Figure 4.4B). However, the similarity ratings of the face recognition algorithm for the two-face average to both encoding images did correlate with participants identification of the average over the new identity, *r*(238) = .19, *p* = .003 (see Figure 4.4C). Thus, in this particular case, the algorithm is consistent with the behavioural data from human participants, who can see the similarity of a two-face average to its corresponding identities.

Figure 4.4

Scatterplots demonstrating the correlation between algorithm similarity scores for the two-face average or exemplar and human accuracy in Chapter 2 Experiment 3 for trials where the exemplar was paired against the average (A), the exemplar was paired against a new face (B) or where the average was paired against a new face (C). Each point represents an individual stimulus.



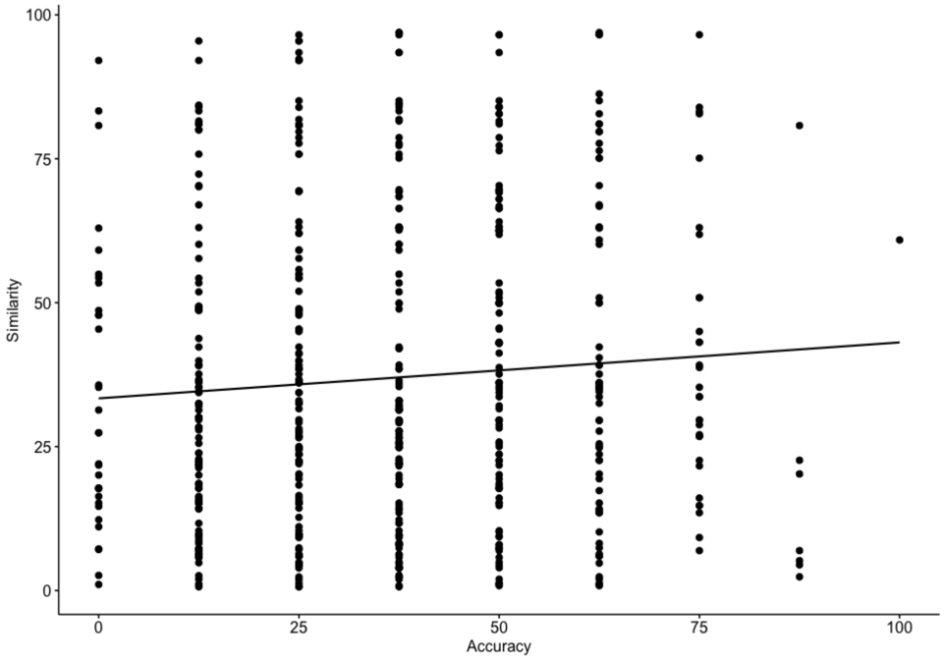
The correlation between the selection frequency of a two-face average over a new identity with the similarity of two-face averages provided some evidence that humans use this similarity in their identification decisions. If this also applies to four-face averages, which contain lower similarity to the encoding images, then this provides preliminary evidence that the similarity between an average and its constituent identities could be a mechanism behind the ensemble coding of identity. The same analysis was therefore repeated for the data of Experiment 4 from Chapter

3, which investigated whether human observers are able to detect the resemblance of a four-face average to a set of encoding images was investigated. This experiment showed that observers have difficulty detecting this resemblance, with a mean accuracy of only 51.6%. Whilst algorithm similarity ratings of the four-face average to its constituent identities was low at an average of 36.3%, this was still higher than for non-matching exemplars, which had an average similarity of 1.5%. It is therefore possible that observers were able to encode this similarity below a decision-making threshold. For example, whilst observers show poor ability to explicitly detect resemblance, the same observers show an implicit bias towards the faces that have the strongest resemblance (Platek et al., 2003). Moreover, the strength of ensemble coding effects depends on the range of items in a set. Sets with a larger variety of items (e.g., eight colours) result in weaker ensemble effects than sets with a smaller range (e.g., two colours) even with the same number of items (Maule & Franklin, 2015). This is because, with larger variability, the average resembles each individual item less. Additionally, similar faces are more likely to be encoded into an average (Neumann et al., 2015). Therefore, higher similarity among set members will result in an average that has greater similarity to each member. Participants may then be recognising the average because this similarity passes the decision-making threshold, whereas for more variable sets, the similarity between the encoded images and the average reaches this similarity less frequently, resulting in weaker ensemble effects.

To explore if participants were encoding the similarity between a four-face average and its constituent identities below a decision-making threshold, the Amazon Rekognition similarity scores for the four-face average were correlated with participant's hit accuracy in Experiment 4 of Chapter 3. Faces rated familiar by participants were not removed for this analysis, as this would have resulted in very few data points per face (*M* = 7.1, *Mode* = 8, *Min* = 3). Average hit accuracy - whereby participants correctly selected a constituent identity - for all items, both familiar and unfamiliar, was 36.2%. However, participants' hit accuracy for each face did not correlate with the Amazon Rekognition similarity of the four-face average, *r*(598) = 0.08, *p* = .063. This provides further evidence that the similarity of a four-face average to its constituent identities is not readily perceivable to observers.

Figure 4.5

Scatterplot showing the correlation between human accuracy in Chapter 3 Experiment 4 and the similarity ratings of the four-face average by Amazon Rekognition. Each point represents an individual stimulus.



4.5. General Discussion

The aim of this chapter was to investigate how much similarity an average has to its constituent identities. This was examined using Amazon Rekognition – a facial recognition algorithm – to gain a more sensitive and consistent measure of similarity than human observers can provide. To first establish the accuracy of Amazon Rekognition, two well-established tests of facial recognition (the GFMT and KFMT) were run through the algorithm (see Burton et al., 2010; Fysh & Bindemann, 2018). Perfect accuracy was achieved on both tests, highlighting the high capability of the algorithm. Similarity ratings and match decisions were then collected between the encoding images and the exemplars, non-matching exemplars, two-face averages, and four-face averages that were used in Chapters 2 and 3. Similarity between the encoding images and the two-face averages was close to Amazon Rekognition's decision threshold (86%), whereby two images are considered an identity match if the similarity score passes this threshold. Many cases (65.8% of face pairings) were classified as a match by the algorithm, demonstrating that two-face averages often contain sufficient similarity to be considered a match by an algorithm. Four-face averages, on the other hand, had much lower similarity ratings to their constituent encoding images, and were very rarely accepted as a match by the algorithm (3.3% of face pairings).

As yet, no studies have directly investigated how much similarity an average has to its constituent identities. In most studies, the formation of an internal average is *assumed*. These studies typically employ a face matching paradigm that is similar to those used in Experiments 1 and 2 of Chapter 2 (see, e.g., Bagaïni & Hole, 2017; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2013; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). Whilst this enables the investigation of whether participants can mistake an average as a member of an encoding set, it does not provide direct evidence that the internal representation is an average. For example, it is possible that the average is simply similar enough to an encoded exemplar to be mistaken as such, especially when the amount of variability is considered that observers have to tolerate across different ambient images of the same identity (see, e.g., Jenkins et al., 2011). Contrary to this reasoning, Chapter 3 demonstrated that observers have trouble in explicitly detecting the constituent identities of an average, particularly when the average was comprised of four identities. However, humans also judge similarity in an inconsistent manner (Fysh & Bindemann, 2023), while visual perception is also influenced by factors such as verbal cues and motivational biases (Balcetis & Dunning, 2006; Lupyan & Spivey, 2010). It is therefore possible that human judgements are dependent on the nature of the task. For instance, judgements of similarity differ when the stimuli are presented on a web page compared to a paper print out (Song, 2013). This does not necessitate that humans cannot *perceive* the similarity, but that they may respond differently due to task demands or expectations (e.g., Jones et al., 2015; Meinhardt et al., 2014). Chapter 3 required participants to make a yes/no decision as to whether each selection identity was a constituent identity of the average. This required a decision threshold, whereby each face would be considered a match if it passed this threshold. It is possible that this does not fully measure whether observers can see any similarity there and if they use this to inform their decisions.

To investigate whether human observers are able to use the similarity of an average to the encoding images in their identity matching judgements, the similarity scores generated by Amazon Rekognition were compared against the human data from Experiment 3 in Chapter 2, where participants were asked to select whether a two-face average or an exemplar most resembled one of two encoded targets, and Experiment 4 in Chapter 3, in which participants were asked to identify which encoding images were used to create a four-face average. Whilst similarity between the two-face average and the encoding images was found to correlate with whether participants would select the average over a new identity in Chapter 2, no correlation was found between the similarity between a four-face average and the encoding images with the data from Chapter 3. Overall, this suggests that most four-face averages do not contain sufficient similarity with the encoding faces for human observers to detect.

On the other hand, two-face averages contained enough similarity to the original encoding images to be detected by both algorithms and humans. Whilst the mean similarity score of the two-face average to the encoding images did not pass the algorithms threshold, this was close to threshold with many cases being accepted as a match. Moreover, the similarity of the exemplar to the encoding image was only found to be higher than the two-face average by 17.4%. At the same time, there was a large range of similarity scores, even for the exemplar faces, with a difference of 70.1% between the highest and lowest similarity rating. Humans therefore have to deal with a large range of similarity for *matching* identities. It is therefore perhaps not surprising that a difference of 17.4% between exemplars and two-face averages can go undetected by observers.

It is clear that a single face can show a large range of variability (e.g., Jenkins et al., 2011). For unfamiliar identities, this can cause difficulties in identity matching, whereby two different but similar faces may be mistaken as a match, but two variable images of the same person are identified as a mismatch (e.g., Fysh & Bindemann, 2023), Accurate processing of this within-person variability is therefore thought to underlie successful face learning (e.g., Baker & Mondloch, 2019; Kramer, Jenkins, Young, et al., 2017; Ritchie & Burton, 2017). For example, exposing participants to more variable images of an identity promotes face learning compared to face sets exhibiting low or no variability (Kramer, Jenkins, Young, et al., 2017; Ritchie & Burton, 2017). However, for unfamiliar faces, the extent of this variability that is required for face learning is unknown. It is possible that a two-face average falls within the limits of the within-person variability around an identity, so that it is similar enough to be accepted as an ambient image of one of the encoding targets. This might also be more likely to occur if the encoding set has not been encoded sufficiently, so that these limits are not established clearly. For instance, many ensemble paradigms present the encoding set only briefly (i.e., 2000 ms; Ji & Hayward, 2021; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018), which may not provide observers with sufficient time to encode the set (e.g., Fahsing et al., 2004; MacLin et al., 2001; Read et al., 1990; Reynolds & Pezdek, 1992). Even with exposure durations of 3000 ms, memory for specific facial features is reduced compared to longer durations (i.e., 20 seconds; Reynolds & Pezdek, 1992). It is therefore possible that under short exposure times, observers encode a less-detailed representation of the face, similar to how observers have been shown to encode scene ‘gist’, by storing a general, less-detailed overview of a scene (Furtak et al., 2022; Rousselet et al., 2005). This could result in a more liberal identification threshold, whereby two-face averages that are most similar to the encoding images to be accepted as a match.

In contrast, four-face averages should retain a much lower level of similarity to the original encoding images, by virtue of being constructed from a larger set of identities. The experiments reported in this thesis demonstrate that both humans and algorithms have difficulty seeing the resulting resemblance of a four-face average to its encoding images. Despite this, many studies on the ensemble coding of identity use encoding sets comprised of four or more target faces (e.g., Bagaïni & Hole, 2017; Matthews et al., 2018; Peng, Kuang, et al., 2019; Sama et al., 2019). The current experiments demonstrate that the identification of an average as a face from the encoding set should be highly unlikely under these conditions if observers are selecting such a probe face if this is based on its visual resemblance to the original encoded identities. This points towards an alternative explanation for such findings based on the formation of an *internal* average from a set of encoding faces. Accordingly, an *external* average is selected at test because it matches an internal average of the encoding identities, rather than because of its limited visual resemblance to the individual encoding faces.

CHAPTER 5:

Summary, Discussion, and Future Research

5.1. Summary and Conclusions

This thesis investigated the role of similarity in the ensemble coding of facial identity. Ensemble coding refers to the process in which observers extract the average of a set of stimuli (Ariely, 2001; Epstein & Emmanouil, 2017; Haberman & Whitney, 2009). This occurs for a wide range of low-level visual features, including size, motion direction, and colour (e.g., Ariely, 2001; Epstein & Emmanouil, 2017; Rajendran et al., 2021; Sweeny et al., 2013; Sweeny et al., 2015). More recent research has found that this phenomenon also occurs for high-level visual features such as faces. For instance, ensemble coding has been found to occur consistently for emotional expressions, whereby an average emotional intensity is rapidly extracted from a set of faces (Haberman & Whitney, 2007; Haberman & Whitney, 2009). This occurs under a wide range of conditions, such as with set sizes ranging from 2 to 24 items and with viewing times of 50 ms to 2500 ms, indicating that this effect is robust (Haberman & Whitney, 2009; Leib et al., 2014; Neumann et al., 2018; Yang et al., 2013).

Research has since begun to investigate whether identity can also be encoded into an ensemble average. In these studies, observers are required to encode a set of faces before being presented with a probe face. Observers are then asked to either adjust the probe along a continuum of morphed faces until it matches the mean identity of the set (i.e., adjustment paradigm; Bai et al., 2015; Leib et al., 2014; Sama et al., 2019) or report whether this probe face was a member of the encoding set (recognition paradigm; Bagaïni & Hole, 2017; Ji & Hayward, 2021; Matthews et al., 2018; Neumann et al., 2013; Neumann et al., 2018; Peng, Kuang, et al., 2019; Peng, Zhang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015; Rhodes et al., 2018; Robson et al., 2018). Whilst the adjustment paradigm demonstrates that observers can accurately judge the set mean, research using the recognition paradigm is particularly interesting as the average is frequently *mistaken* as a member of the encoding set in these paradigms. Previous work using this recognition paradigm has shown that the average of a set of faces can be mistaken as often, and sometimes more frequently, than a specific member taken directly from the encoding set (Bagaïni & Hole, 2017; Matthews et al., 2018; Peng, Kuang, et al., 2019; Peng et al., 2021; Peng et al., 2022; Rhodes et al., 2015).

However, these ensemble effects for identity also appear to differ from those observed for other visual stimuli. For both low-level features, such as object size, and some high-level features, such as emotion and lifelikeness, where observers have to report how alive the items in the set appear, the average is accurately recognised but the individual set members are not (e.g., Brady & Alvarez, 2011; Haberman & Whitney, 2009; Leib et al., 2016). This suggests that ensemble coding is a mechanism to rapidly encode a *summary* of visual information when there are insufficient cognitive resources to encode *all* information. However, in identity ensemble paradigms, both the average and exemplars from the encoding set are often recognised (e.g., Bagaïni & Hole, 2017; Neumann et al., 2013; Neumann et al., 2018; Sama et al., 2019), which contradicts the notion that ensemble coding arises from some kind of capacity limits. An alternative explanation might be that an average face is selected in ensemble-coding recognition-paradigms because it bears high similarity to one or several encoded exemplars.

One possible method for investigating the internal facial representations that are formed during ensemble-coding recognition-paradigms is to compare which stimulus – the average or its constituent exemplars - is selected more frequently. In ensemble paradigms, this would be achieved by measuring which face – the exemplar or the average – is selected more frequently at recognition. However, the existing evidence is mixed as to which representation is selected more frequently (Bagaïni & Hole, 2017; Leib et al., 2014; Matthews et al., 2018; Peng, Kuang, et al., 2019, Sama et al., 2019). This leaves the possibility that the average might be recognised because it *resembles* the encoded exemplar. Whilst there is always likely to be *some* similarity between the average and the exemplar, the extent to which this can be detected and utilised when viewing these faces is important in determining whether identification of the average reflects ensemble coding. This was investigated throughout this thesis through a series of seven experiments.

Chapter 1 examined the relative strength of exemplars and averages by investigating which of these faces would be selected more frequently in an ensemble recognition paradigm. Many previous studies using this paradigm use complex designs, with large set sizes ranging between 4 and 18 identities (e.g., Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Robson et al., 2018). With larger set sizes, it is possible that not all identities are encoded sufficiently due to factors such as attention and capacity limits in encoding (Bindemann et al., 2005; Guo et al., 2009; Leder et al., 2010). In this case, the average might have an advantage over the exemplar. For example, if a probe face was an exemplar which was not initially encoded due to capacity limits, then recognition at test should be low. In contrast, the average would contain similarity at least to the identities that *were* encoded from the set and therefore have a general advantage over the exemplars at test.

To examine the relative strength of exemplars and averages, this study employed a recognition task, whereby observers were shown an encoding set before being presented with a probe face and asked to report whether this face was from the encoding set or not. To overcome the complications with the large set sizes typically used in ensemble paradigms (e.g., Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Robson et al., 2018), this study reduced the encoding set to a logical minimum of two identities. This provided optimal conditions for the encoding of all targets and therefore reduced the possibility of observers only encoding a subset of the set, enabling a more direct comparison between recognition of the exemplar and the average. An additional condition was also included, whereby observers were only shown a single identity at encoding. This provided an important novel comparison. If the average is recognised because it resembles an encoded exemplar, recognition of the average should be equivalent across one- and two-target encoding sets as the similarity between the exemplar and average would remain constant. However, if the average is recognised because it has been internally encoded, then the average should not be recognised when only one target is encoded.

Experiment 1 used the same images at encoding and test. The results demonstrated that the average was selected as often as the exemplar, providing initial evidence for the encoding of both an average *and* its exemplars. Importantly, this experiment also showed lower recognition frequency of the average in one-target than two-target conditions. This provided the first study to directly compare recognition of the average when multiple targets are encoded compared to a single target, enabling the potential selection of the average based on its similarity to the exemplar to be investigated. As the average was selected less frequently in the single-target condition in this experiment, this lends support to ensemble coding.

To compare whether these effects reflect the encoding of identity or simpler image-based effects, Experiment 2 replicated Experiment 1 but with different images of the same identities at encoding and test. By using different images at encoding and recognition in Experiment 2, this provided evidence that these ensemble effects were indeed for identity, rather than simpler image-based effects, as ensemble coding was found to occur across a change in image. This is an important distinction, as it shows ensemble coding of identity is not a product of ensemble coding of the lower-level elements in an image, such as colour and shape, but its own independent effect. Again, the exemplar and average were found to be selected at a comparable rate.

Finally, to assess the strength of these representations in a more direct way, Experiment 3 presented both the average and the exemplar simultaneously at recognition to put these faces into direct competition. Participants were then asked to select which of these faces most resembled one of the two encoded identities. The logic behind this was that the face that is selected more frequently would likely be the stronger representation, as this would be a direct match to the internal representation. However, under these conditions, both the average and the exemplar faces were selected at a comparable rate. This provided further evidence for the encoding of *both* an exemplar and an average.

Overall, this chapter provided preliminary evidence that the average is selected because it is encoded alongside an exemplar, rather than because of its similarity to the encoded exemplars. However, this does not completely rule out the possibility that similarity can explain these effects. Consider a scenario where both encoding identities are sufficiently encoded and stored as an exemplar. If only the exemplars have been encoded and not an average, whilst the average face will still have some similarity to both encoded exemplars, the exemplar images will be a direct match to what is stored internally and therefore should be selected more often. On the other hand, if the exemplars are encoded to a limited extent, then the encoded representation might no longer be a direct match to the external exemplar. Therefore, recognition of the exemplar depends on the similarity to what has been encoded and the external image. Whilst this similarity should be high, it is also possible that the level of familiarity that an observer has to the exemplar is at a level similar to the average face, especially as the average face would contain similarity to both encoding identities, which could boost overall familiarity.

However, this argument is built on the assumption that observers are able to see the similarity between an average and its constituent identities. Research on facial morph detection, where two faces are averaged together by varying amounts (i.e., 20:80, 50:50, 30:70) and observers are asked to report whether the face is a morph or an ambient image of an identity, shows that the less a second identity contributes to the morph (i.e., 20% in a 20:80 morph), the less likely it is that the morph will be detected by observers (Rotshtein et al., 2005). Ensemble paradigms typically use set sizes of four or more faces (Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Robson et al., 2018), whereby each identity contributes 25% or less to the average. This leaves the question of whether observers can see the similarity between a four-face average and its constituent identities in the first place.

Consequently, Chapter 3 investigated the extent to which observers are able to see the similarity of an average to its constituent identities. In Experiment 4, participants were provided with a four-face average in the centre of the screen surrounded by four selection identities. Participants were asked to select which of these identities, if any, were used to create the average. This introduced a new approach as participants were provided with an average and asked to decode the identities that were used to create it. This is in direct contrast to paradigms typically used to investigate ensemble coding, whereby the encoding set is shown *followed* by a probe face. In these paradigms, selection of the average is assumed to represent ensemble coding, overlooking the possibility that the average is selected because it resembles an encoded exemplar. By presenting participants with the average and asking them to select the identities used to create this, this study enabled an investigation into how much similarity can actually be seen directly between an average and its encoding identities. This provided observers with unlimited comparison time and removed any memory demands and therefore provided the optimal conditions to detect similarity. In further experiments with this paradigm, the average was also presented *before* its constituent identities. This ensured that the encoded face was an average and therefore investigated observers' abilities to see the similarity between an encoded average and its constituent identities.

Even under the ideal conditions to detect similarity, where observers had unlimited comparison time, performance on this task was low at 51.6%, highlighting the difficulty of this task. In a four-face average, each identity contributes 25% to the average. This might not be a sufficient amount of identity for observers to detect. In this task, there was no encoding of internal average. Instead, participants were provided with the average and were able to directly compare each selection identity against the average. Despite this, performance was at chance, with half of the identities used to create the average missed by participants and non-matching identities accepted as a match over 25% of the time.

Experiment 5 aimed to investigate whether the low accuracy found in Experiment 4 was due to the amount of identity information retained in a four-face average or whether this reflected difficulty in the task itself. To do this, two conditions were added – a two-face average and four exemplar images as the target. In a two-face average, each identity contributes 50%, and therefore, performance was expected to be higher for the two-face average than the four-face average if accuracy on this task was related to the amount of identity information that can be gleamed about an exemplar identity from an average. In contrast, in the four exemplar image condition, four images of each selection identity were included in the centre of the screen. These were different images of the same identities as those that were used as the selection identities. By including this condition, this experiment was able to provide a baseline level of accuracy, as it is clear that matching two different images of an unfamiliar identity can be challenging (e.g., Burton et al., 2010; Megreya & Burton, 2006a). Therefore, by including four exemplar images for participants to match against different images, this condition established accuracy when observers are required to match four individual unfamiliar faces.

The results demonstrated that observers were better at detecting the similarity between two-face averages and its constituent exemplar identities than with the four-face averages. Accuracy was also higher for the four exemplar images than the four-face averages. This provides evidence that the lower performance for the four-face average can be attributed to the reduced identity information from each target in the average, as two-face averages, which include more information from each target, and different images of the target resulted in higher accuracy. Interestingly, accuracy for the two-face average was at a level comparable to the four exemplar images, suggesting perhaps that two-face averages contain similar levels of similarity to each identity as a different image of that identity.

While these experiments so far utilised ensemble-coding matching-paradigms, an important component of previous ensemble-coding studies is *memory*, whereby observers have to compare the similarity of the probe face against its internal representation. As yet, no studies have asked observers to encode an *average*. This is an important manipulation because it ensures the internal representation is an average. Therefore, the paradigm used in Experiments 4 and 5 was modified so that the average (two- or four-face) or four exemplar images were provided at an initial encoding phase in Experiment 6. Following this, participants were then presented with the four selection identities without the probe stimulus in the centre. This study examined the extent to which observers can detect the similarity of individual exemplars to an internal representation. However, accuracy both four-face averages and two-face averages remained similar to the levels seen in Experiments 4 and 5, suggesting the similarity between an encoded average and an external average to its constituent identities remains stable with and without a memory demand. On the other hand, performance for the four exemplar images was found to be at a comparable level to the four-face average when a memory demand was introduced. One potential reason for this is that when encoding a set of four faces, observers encode the average, and this condition therefore becomes equivalent to the four-face average condition. In contrast, in the four exemplar image condition, observers have to encode *four* faces as opposed to a *single* average face. Therefore, performance may have been at a lower level in the four-face condition compared to Experiments 4 and 5, because this condition becomes comparatively more challenging with a memory demand.

To explore this possibility, Experiment 7 included a working memory task in-between encoding and selection. This involved viewing a set of objects before participants were asked if a probe object was from the set or not. This memory task included both a high-load condition, whereby observers were given a set of four objects to view, and a low-load condition, wherein the set was comprised of only two objects. If the reduced performance of the four exemplar images in Experiment 6 could be attributed to the memory demand of having to encode four faces compared to a single average, then the high-load condition should have selectively impaired identification of the four exemplar images. However, if participants formed an average of the four exemplar images at encoding, then the working memory load manipulation should not have selectively impaired performance on the four exemplar image condition relative to the other conditions.

In Experiment 7, performance on the object memory task was poorer on high-load than low-load trials, showing that this manipulation was effective. Despite this, no effect of working memory load on performance in the selection phase of the resemblance task was found. This could provide evidence for ensemble coding of identity as it is possible that the four exemplar images were being combined into an average representation, resulting in comparable performance in four-face average trials and four exemplar image trials. However, it is also possible that the object memory task did not interfere with the encoding of the identities. For instance, previous research has suggested that working memory load might be specialised (e.g., Cheung & Gauthier, 2010; Park et al., 2007). These studies demonstrate that memory for faces is only affected by working memory load if the working memory task also involves faces, but not objects (Cheung & Gauthier, 2010; Park et al., 2007). Therefore, the object memory task may not have been sufficient to induce a working memory effect on the resemblance task with faces.

Overall, Chapter 3 demonstrates that the detection of similarity between an average and its constituent identities is difficult and, in doing so, provides further evidence for ensemble coding. However, whilst it seems that observers have difficulty detecting the similarity between an average and its constituent identities, similarity judgements made by humans can be inconsistent both within- and between-observers (see, e.g., Alenezi et al., 2015; Bindemann & Sandford, 2011;

Bindemann, Avetisyan, et al., 2012; Russ et al., 2018). Moreover, human responses can be affected by the type of task used and often exhibit response biases, whereby they are likely to respond in a certain way (i.e., frequently selecting two faces), particularly when they are uncertain about the correct solution for a task (e.g., Aminoff et al., 2012; Kantner & Lindsay, 2012).

Human similarity judgements can also be influenced by context. For instance, in one study, participants gave similarity ratings for pairs of faces. These faces were then presented in another phase alongside a family name. After this phase, participants re-rated the similarity between the face pairs. Findings showed that similarity ratings for face pairs encoded with different family names decreased but increased for those encoded with the same family name (Ashby et al., 2020). Whilst both human and algorithm accuracy are still dependent on factors related to the stimuli, such as the subject's gender or age, the angle of the face, and image characteristics (e.g., lighting, resolution; see Givens et al., 2013), algorithms follow a sequence of computations which consistently produce the same output. Therefore, whilst human accuracy can be inconsistent, algorithms provide a more consistent and reliable measure of similarity. Because of this, the similarity between an average and its constituent identities was further explored in Chapter 4 using a more accurate and consistent measure than human observers by employing a facial recognition algorithm.

These algorithms can now surpass human observers in their ability to detect identity (e.g., Carragher & Hancock, 2023). In Chapter 4, the Amazon Rekognition algorithm was used to generate similarity scores between the encoding images and the averages and exemplars. In a first step, the face-matching capabilities of this algorithm were demonstrated with the GFMT and KFMT by running the face pairs through the algorithm to determine if these were classified as a match or mismatch by the algorithm (Burton et al., 2010; Fysh & Bindemann, 2018). Then, similarity scores were generated for the encoding images with exemplars, two-face averages, four-face averages, and non-matching exemplars. On average, exemplars were given a high similarity score of approximately 99% to the encoding images with ~99% of these being classified as a match compared to non-matching exemplars, which were never accepted as a match and were given a score of 1.5%. Similarity scores for the two-face averages against the encoding images were also high, with an average score of around 82% and 66% considered matches. On the other hand, similarity scores for four-face averages against the encoding images were only around 36% with ~3% of these considered matches.

Whilst Amazon Rekognition indicates that four-face averages contain little similarity to their constituent identities, two-face averages contain much higher similarity, with over half of the items used passing the threshold to be considered an identity match. These similarity scores between the constituent identities and the two-face average were found to correlate with the human data from Experiment 3 in Chapter 2 in trials where the average was paired against a new identity, but not when the average was directly compared with the exemplar. Additionally, no correlation was found between the similarity between a four-face average and its constituent identities with the human data from Experiment 4 in Chapter 3. As many studies typically use sets of four or more identities (Bagaïni & Hole, 2017; Bai et al., 2015; Leib et al., 2014; Robson et al., 2018), this provides strong evidence that averages retain little similarity to their constituent identities. In turn, this suggests that an internal average must be formed alongside an exemplar in ensemble coding paradigms, as an average should not be selected otherwise, based on its direct visual resemblance to faces from an encoding phase.

Throughout this thesis, face exemplars and averages were recognised at a comparable rate across a range of paradigms and methodologies. This corresponds to previous findings, where the exemplar and average are often selected as frequently as each other (Bagaïni & Hole, 2017; Neumann et al., 2013; Neumann et al., 2018; Sama et al., 2019). However, previously, no research had explored the role of similarity in ensemble coding. This thesis attempted to fill this gap in knowledge. The experiments here demonstrate that whilst two-face averages contain enough similarity to their constituent identities to be identified, four-face averages often do not. This creates a paradox, whereby if the average can be misidentified as a member of the set, it should resemble the set members. Yet, this thesis demonstrates that the similarity between a four-face average and its constituent identities is remarkably difficult to detect.

This points to exemplar coding and ensemble coding as independent and parallel processes. Previously, there has been doubt on whether exemplar encoding is a requirement for ensemble coding to occur or if these reflect separate, independent processes. Some research suggests that there is no evidence of ensemble coding of identity without the encoding of individual exemplars (Neumann et al., 2018). If the recognition of averages required the encoding of exemplars, then it would be expected that observers would be able to see the similarity between an exemplar and average. However, the findings from Chapter 4 show that observers struggle to see this similarity. There is indication of reduced exemplar recognition with an increase in encoding set size, but no reduction in average recognition (Li et al., 2016), providing evidence for distinct processes. Additionally, patients with prosopagnosia show an inability to recognise individual faces but can accurately encode the average identity and emotion of a set of faces (Leib et al., 2012; Robson et al., 2018). Considering these findings alongside the results of Chapter 4, this suggests that exemplar and ensemble coding are underpinned by distinct processes.

Despite a now extensive body of research, very little is still understood about the mechanisms behind ensemble coding of identity. One mechanism might relate to how the visual system deals with variability in appearance. The face of a person can show substantial variation in appearance, with even superficial changes in viewpoint or hairstyle changes having large effects on an observer's ability to recognise the identity (Bartel et al., 2018; Stephan & Caine, 2007). Despite this, familiar faces can be recognised easily (e.g., Bruce et al., 2001; Lander & Bruce, 2001). One theory is that faces are represented as an average identity, whereby each encounter with a face is incorporated into an average for that identity (Burton et al., 2005). During the construction of these averages, image-specific variance is filtered out, retaining only information diagnostic of identity. However, for an ambient image to be recognised using this average template, some variability must be accepted to enable recognition despite image-specific characteristics. Therefore, if this is the mechanism ensemble coding draws upon, it is perhaps not surprising that exemplar information is also encoded at some level.

Related to this is the finding that for low-level visual features, additional summary statistics can also be extracted. For instance, one study investigated ensemble coding of three different visual dimensions, comprising of size, orientation, and brightness (Khayat & Hochstein, 2018). In this study, they presented participants with sets of 12 stimuli varying along one of the three dimensions. Participants were then presented with two items and asked to select which of these items was a set member. These items were either a member of the set, a non-member, or the mean of the set. Participants were more likely to endorse the set mean, but also a set member the closer it fell to the mean of the set. Non-members were also more likely to be selected the closer they fell to the mean, and performance was at chance if the non-member was closer to the mean than the actual set member. However, if the non-member fell outside of the range of the set, then accuracy was much higher (Khayat & Hochstein, 2018). The same pattern is also found in emotion ensemble coding studies, whereby observers are perceptive to the range of the set (Haberman & Whitney, 2009). If the range of low-level visual features can be extracted, then it is also possible that the range can be extracted for higher-level features such as identity.

This is logical if the mechanisms behind the ensemble coding of identity draw upon similar mechanisms to face learning. For example, familiar faces can be recognised without much difficulty despite large variations in appearance (e.g., Bruce et al., 2001; Jenkins et al., 2011). One theory that suggests how this happens is the “islands of expertise” theory (Hancock, 2021). Accordingly, as humans become familiar with a face, they learn how much that face can vary. This creates an “island” where any instances that fall within this island are accepted as a match for that identity. Some islands may overlap. However, with increasing familiarity, these islands become more distinct. Unfamiliar faces, on the other hand, have no existing island and observers therefore have to resort to feature matching. For instance, observers are better at remembering a specific image they have seen of an unfamiliar face than they are a familiar face, but are more accurate at recalling the identity for familiar faces (Armann et al., 2016). However, if an unfamiliar face resembles a familiar face, they can also activate a familiar face’s island if they fall within the range of variation. This can be strong enough to cause misidentifications, but can also be used to boost future familiarity with the unfamiliar face.

One question this raises is how these islands are activated. Literature on object recognition has proposed a hierarchical processing of objects that enable accurate categorisation of the object (e.g., Ayzenberg & Lourenco, 2019; Large & McMullen, 2006; Sanocki, 1993). These start with low-level properties of the objects, such as their global shape, and continue along to higher-level properties, such as more specific features of the object (e.g., intricate details, colour, texture). This also applies to letter recognition. For instance, letters are recognised across a wide variety of fonts (e.g., Polk et al., 2009). This is thought to happen through an initial encoding of the features of the letters, such as horizontals and terminations (Finkbeiner & Coltheart, 2009; Fiset et al., 2009). These enable the activation of an abstract representation which enable the recognition of the letter regardless of the font (Finkbeiner & Coltheart, 2009). However, with more abstract fonts which deviate from the typical structure of a letter, the initial stage is disrupted, resulting in lower accuracy or slower processing of the letters and more reliance on contextual information (Oderkerk & Beier, 2024).

For identity, a similar mechanism might be involved. It has been proposed that there are facial recognition units (FRUs), which contain structural information about familiar faces (Bruce & Young, 1986). When encountering a familiar face, structural properties of the face are encoded which activate corresponding FRUs. The strength of the FRU activation is dependent on how well the encoded structural properties of the face match the stored representation. However, to ensure that the stored representation is activated by the many different instances of a known face that might be encountered in everyday life, these need to accept some variability in a person’s appearance as that identity. It is possible that the representation is an average, whereby the average can shift with new experiences of that identity (e.g., Burton et al., 2005). To allow for variability, there must also be a range around this average where each new encounter can then be accepted as a match to that representation if it falls within this range. For familiar faces, this range is already known, and observers can accurately separate within-person variability from between-person similarity. However, for unfamiliar faces whose range of variability is not known, high within-person variability can lead to a single identity being mistaken as several different individuals (Jenkins et al., 2011).

A related question is how these representations are formed. As unfamiliar faces have no pre-existing representation or “island”, each new encounter needs to be matched to the unfamiliar face to start forming these representations. It is not clear how different instances of a person are cohered into a single representation. One possibility is that this is regulated by context. For instance, in face matching tasks, observers are more likely to accept two images of different people as a match if the names of the two images match (Trinh et al., 2022). Additionally, two images are more likely to be correctly matched if learned under the same name compared to two different names (Menon et al., 2018). Another way faces might be combined into a single representation is through co-occurrence. For instance, when observers encode a pair of faces, one of these faces can result in priming effects for the second face (Vladeanu et al., 2006). This demonstrates that simple co-occurrence can facilitate integration of visual information. It is possible that, when co-occurring as a group, the visual system integrates the identities and automatically extracts the average and the range, enabling the system to form an island of the set.

Alternatively, as observers have no representation of the unfamiliar faces, they therefore have no information about the variability that such a face can demonstrate. According to the islands of expertise theory (Hancock, 2021), unfamiliar faces are therefore matched based on features that can be image dependent. It is possible that the similarity of the average to the pictorial representation might be sufficient to activate the representation. Additionally, if each face is activated in a similar way by viewing the average, these activations may combine to facilitate a sense of familiarity with the average, even if each individual activation is only weak. However, the findings of this thesis show that the average is selected at a comparable rate to the exemplar, even when paired directly against the exemplar, suggesting that similarity in itself is not sufficient to explain why the average is selected.

These findings provide firsthand evidence that averages retain very little similarity to their constituent exemplars. This points to an internal representation that *is* an average. The formation of an average of a set of unfamiliar faces provides valuable insight into possible mechanisms of face learning. It seems likely that encoding the average and range enables recognition of a face across within-person variation in appearance. However, as different encounters with unfamiliar identities need to be combined to form the representation to begin with, how these are matched without this representation remains unanswered. It is possible that ensemble coding is a by-product of the cognitive system's attempt to form this representation, as faces encoded together may be linked by the visual system through context and simple co-occurrence.

5.2. Future Directions

As the experiments in this thesis show that ensemble coding effects cannot be explained by similarity, this raises several avenues for future research. For instance, in Chapter 3, human perception of the similarity between an average and its constituent identities was measured using a matching task, where observers had to select which faces were used to create the average. In this paradigm, observers then had to make a yes/no judgement for each of the selection identities. This requires a decision threshold, whereby faces that are perceived to pass a certain similarity threshold are considered a match. However, a face that falls just below this threshold may still be recognised by an observer as having some similarity to the average. Chapter 4 attempted to investigate this further by correlating human performance with similarity ratings generated by a facial recognition algorithm, finding no correlations between the similarity of a four-face average to its constituent identities. However, it is clear that human perception of similarity is complex (e.g., Alezeni et al., 2015; Bindemann, Avetisyan, et al., 2012; Bindemann & Sandford, 2011; Fysh & Bindemann, 2023; Russ et al., 2018) and may not be adequately represented by algorithm similarity scores. For instance, the same person will make different similarity judgements to the same items at different times (Bindemann, Avetisyan, et al., 2012). It is therefore possible that a yes/no decision did not capture the full role of human similarity perception. One way to investigate this further would be to ask participants to make graded similarity judgements as opposed to such binary decisions. This could provide a more sensitive measure of similarity.

Additionally, it is possible that similarity and resemblance reflect distinct cognitive processes, yet so far, no distinction has been drawn between these definitions. One possible way to draw a distinction between these is to consider similarity as referring to a judgement made between two *external* items. Similarity judgements can therefore be made for two items by any observer, irrespective of whether they are familiar or unfamiliar with an identity. Resemblance, on the other hand, could be used to refer to an external stimulus for which an internal cognitive representation also exists. This is more subjective than similarity, as it depends on the individual's internal representation of the identity and requires familiarity with the face as a prior internal representation needs to have been formed. By this logic, both similarity and resemblance were measured in this thesis. For example, in the experiments in Chapter 2 resemblance was investigated primarily, as participants were comparing the probe faces with their internal representations. Resemblance was also measured in Experiments 6 and 7 in Chapter 3 when participants were required to encode the average before selecting the constituent identities. In contrast, Experiments 4 and 5 in Chapter 3 presented an external average alongside the selection identities, and therefore measured similarity. This potential distinction between similarity and resemblance could be investigated further by comparing resemblance ratings to similarity ratings. If a distinction between the two processes is found, then this provides a valuable framework for research investigating the cognitive representation of faces and other visual phenomena, allowing a more direct measure of the internal representation.

For instance, this could be used to further investigate the internal representation of a set of faces in ensemble coding studies. Similarity ratings, where both an encoding image and a probe face (an exemplar or an average) are shown simultaneously after viewing an initial encoding set, could be compared to resemblance ratings, where the encoding set is presented prior to a single exemplar or average probe. If the internal representation is an average, then one possible outcome could be higher resemblance ratings for the average than the similarity ratings. This is because resemblance should be high between an internal and external average. However, in the similarity task, the average is compared against a face that only contains half of the identity information in the average, and therefore the similarity should be lower. Yet, if the internal representation is an exemplar, then the average is no longer a direct match to the internal representation. Therefore, it is likely the similarity ratings for the average would be higher than its resemblance ratings, as previous work has shown that similarity is easier to detect in matching tasks than recognition tasks (e.g., Bobak et al., 2016; Estudillo & Bindemann, 2014).

From the similarity scores in Chapter 4, it is clear that each constituent identity does not retain the same level of similarity to an average. For instance, a single constituent identity of a four-face average can have a similarity rating to its constituent exemplars of around 80%, yet another constituent identity of the same average may only have a 20% similarity rating. This opens up another interesting avenue for future research. For instance, the question arises of whether higher similarity to the average results in more frequent selection of that exemplar, which would provide stronger evidence for the formation of an internal average. Moreover, what makes certain faces more similar to the average could provide important information about how identity is processed.

Finally, ensemble coding may also explain other phenomena in the literature. For instance, the two-head disadvantage in face recognition refers to the drop in accuracy when participants are required to encode two faces instead of one in eyewitness scenarios. With only one face to encode, person identification from lineups is at ~60% accuracy (Megreya & Burton, 2006a). In contrast, with an additional face to encode, performance falls to ~30% accuracy (Megreya & Burton, 2006a). It is possible that ensemble coding may account for this effect, where the average is encoded from the two faces, which then results in reduced lineup performance when this is tested with an exemplar. Alternatively, ensemble coding could also provide a potential solution to this problem, whereby if observers can extract the average when they are unable to encode the exemplars, perhaps providing an average at test can help overcome this disadvantage.

In conclusion, the ensemble coding of identity appears to be a robust phenomenon that occurs alongside the encoding of individual identities. The experiments reported in this thesis demonstrate that the average is selected at a comparable rate to the exemplar, even when these faces are put into direct competition (Experiments 1-3). However, the similarity between an average and its constituent identities is difficult to detect (Experiments 4-7), suggesting that the comparable selection of exemplars and averages cannot be explained by the similarity of an average to its constituent exemplars. Moreover, even high-performing algorithms struggle to see the similarity between an average and its constituent identities when the average is made of four faces (Chapter 4). Taken together, the experiments in this thesis suggest that an average is not recognised because it resembles an internal exemplar. Rather, the average must be selected because it resembles an internal average. This thesis provides novel insight into the role of similarity between exemplars and averages in the ensemble coding of identity and provides new avenues for future research, in which the mechanisms behind ensemble coding could be further explored.

References

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, 3:e1184. https://doi.org/10.7717/peerj.1184

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392-398. https://doi.org/10.1111/j.1467-9280.2008.02098.x

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345-7350. https://doi.org/10.1073/pnas.0808981106

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., Grafton, S. T., & Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016-1030. https://doi.org/10.3758/s13421-012-0204-6

Andersen, R. A. (1997). Neural mechanisms of visual motion perception in primates. *Neuron*, *18*(6), 865-872. https://doi.org/10.1016/s0896-6273(00)80326-8.

Andrews, S., Jenkins. R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. https://doi.org/10.1080/17470218.2014.1003949

Arantes, J., & Berg, M. E. (2012). Kinship recognition by unrelated observers depends on implicit and explicit cognition. *Evolutionary Psychology*, *10*(2), 210-224. https://doi.org/10.1177/147470491201000204

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157-162. https://doi.org/10.1111/1467-9280.00327

Armann, R. G. M., Jenkins, R., & Burton, A. M. (2016). A familiarity disadvantage for remembering specific images of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 571-580. https://doi.org/10.1037/xhp0000174

Ashby, S. R., Bowman, C. R., & Zeithamova, D. (2020). Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task. *Psychonomic Bulletin & Review*, *27*(4), 791-800. https://doi.org/10.3758/s13423-020-01754-3

Ayzenberg, V., & Lorenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), Article 9359. https://doi.org/10.1038/s41598-019-45268-y

Bagaïni, A., & Hole, G. (2017). Effect of vertical stretching on the extraction of mean identity from faces. *Perception*, *46*(9), 1048-1061. https://doi.org/10.1177/0301006617701160

Bai, Y., Leib, A. Y., Puri, A. M., Whitney, D., & Peng, K. (2015). Gender differences in crowd perception. *Frontiers in Psychology*, *6*, Article 1300. https://doi.org/10.33a89/fpsyg.2015.01300

Baker, K. A., & Mondloch, C. J. (2019). Two sides of face learning: Improving between-identity discrimination while tolerating more within-person variability in appearance. *Perception*, *48*(11), 1124-1145. https://doi.org/10.1177/0301006619867862

Baker, K. A., Mondloch, C. J., & Bindemann, M. (2024). A criterion-placement theory of face matching. https://dx.doi.org/10.2139/ssrn.4947575

Balas, B., Sandford, A., & Ritchie, K. (2023). Not the norm: Face likeness is not the same as similarity to familiar face prototypes. *i-Perception*, *14*(3), Article 20416695231171355. https://doi.org/10.1177/20416695231171355

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, *91*(4), 612-625. https://doi.org/10.1037/0022-3514.91.4.612

Bartel, S. J., Toews, K., Gronhovd, L., & Prime, S. L. (2018). "Do I know you?" Altering hairstyle affects facial recognition. *Visual Cognition*, *26*(3), 149-155. https://doi.org/10.1080/13506285.2017.1394412

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science*, *18*(8), 706-712. https://doi.org/10.1111/j.1467-9280.2007.01964.x

Bicego, M., & Grosso, E. (2019). On the importance of local and global analysis in the judgement of similarity and dissimilarity of faces. *Image and Vision Computing*, *92*, Article 103813. https://doi.org/10.1016/j.imavis.2019.09.004

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277-291. https://doi.org/10.1037/a0029635

Bindemann, M., Burton, A. M., & Jenkins, R. (2005). Capacity limits for face processing. *Cognition*, *98*(2), 177-197. https://doi.org/10.1016/j.cognition.2004.11.004

Bindemann, M., Burton, A. M., Langton, S. R. H., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of Vision*, *7*(10), 1-8. https://doi.org/10.1167/7.10.15

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625-627. https://doi.org/10.1068/p7008

Bindemann, M., Sandford, A., Gillatt, K., Avetisyan, M., & Megreya, A. M. (2012). Recognising faces seen alone or with others: Why are two heads worse than one? *Perception*, *41*(4), 415-435. https://doi.org/10.1068/p6922

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81-91. https://doi.org/10.1002/acp.3170

Bola, M., Paź, M., Doradzińska, Ł., & Nowicka, A. (2021). The self-face captures attention without consciousness: Evidence from the N2pc ERP component analysis. *Psychophysiology*, *58*(4), Article e13759. https://doi.org/10.1111/psyp.13759

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384-392. https://doi.org/10.1177/0956797610397956

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160-1176. https://doi.org/10.1037/xhp0000399

Braje, W. L. (2003). Illumination encoding in face recognition: Effect of position shift. *Journal of Vision*, *3*(2), 161-170. https://doi.org/10.1167/3.2.4

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105-116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Bruce, V. (1994). Stability from variation: The case of face recognition in the M.D. Vernon Memorial Lecture. *Quarterly Journal of Experimental Psychology*, *47*(1), 5-28. https://doi.org/10.1080/14640749408401141

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360. https://doi.org/10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207-218. https://doi.org/10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. W., (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Brunas, J., Young, A W., & Ellis, A. W. (1990). Repetition priming from incomplete faces: Evidence for part to whole completion. *British Journal of Psychology*, *81*(1), 43-56. https://doi.org/10.1111/j.2044-8295.1990.tb02344.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental*

Psychology, *66*(8), 1667-1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223. https://doi.org/10.1111/cogs.12231

Burton, A. M., White, D., & McNiell, A. (2010). The Glasgow face matching test. *Behaviour Research Methods*, *42*(1), 286-291. https://doi.org/10.3758/brm.42.1.286

Carragher, D. J., & Hancock, P. J. B. (2023). Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General*, *152*(5), 1286-1304. https://doi.org/10.1037/xge0001310

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660-675. https://doi.org/10.1037/0096-1523.34.3.660

Chang, A., Murray, E., & Yassa, M. A. (2015). Mnemonic discrimination of similar face stimuli and a potential mechanism for the “other race” effect. *Behavioral Neuroscience*, *129*(5), 666-672. https://doi.org/10.1037/bne0000090

Chang, T.-Y., & Gauthier, I. (2022). Domain-general ability underlies complex object ensemble processing. *Journal of Experimental Psychology: General*, *151*(4), 966-972. https://doi.org/10.1037/xge0001110

Chen, B., & Zhou, G. (2018). Attentional modulation of hierarchical ensemble coding of identities of moving faces. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(10), 1542-1556. https://doi.org/10.1037/xhp0000549

Cheung, O. S., & Gauthier, I. (2010). Selective interference on the holistic processing of faces in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 448-461. https://doi.org/10.1037/a0016471

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393-404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891-900. https://doi.org/10.1016/j.visres.2004.10.004

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition*, *26*(4), 780-790. https://doi.org/10.3758/BF03211397

Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology*, *66*(3), 364-370. https://doi.org/10.1037/0021-9010.66.3.364

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

Codispoti, M., Mazzetti, M., & Bradley, M. M. (2009). Unmasking emotion: Exposure duration and emotional engagement. *Psychophysiology*, *46*(4), 731-738. https://doi.org/10.1111/j.1469-8986.2009.00804.x

Corpuz, R. L., & Oriet, C. (2022). Within-person variability contributes to more durable learning of faces. *Canadian Journal of Experimental Psychology*, *76*(4), 270-282. https://doi.org/10.1037/cep0000282

Cowan, N., Morey, C. C., Chen, Z., & Bunting, M. (2007). What do estimates of working memory capacity tell us? In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The Cognitive Neuroscience of Working Memory* (pp. 43-58). Oxford University Press. 10.1093/acprof:oso/9780198570394.003.0003

Dal Martello, M. F., & Maloney, L. T. (2006). Where are kin recognition signals in the human face? *Journal of Vision*, *6*(12), 1356-1366. https://doi.org/10.1167/6.12.2

Davis, E. E., Matthews, C. M., & Mondloch, C. J. (2024). Ensemble coding of facial identity is robust, but may not contribute to face learning. *Cognition*, *243*, 1-12. https://doi.org/10.1016/j.cognition.2023.105668

DeGutis, J., Mercado, R. J., Wilmer, J., & Rosenblatt, A. (2013). Individual differences in holistic processing predict the own-race advantage in recognition memory. *PLoS ONE*, *8*(4), e58253. https://doi.org/10.1371/journal.pone.0058253

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science*, *28*(2), 193-203. https://doi.org/10.1177/0956797616678188

Elias, E., & Sweeny, T. D. (2020). Integration and segmentation conflict during ensemble coding of shape. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(6), 593-609. https://doi.org/10.1037/xhp0000733

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. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *42*(3-A). 495-512. https://doi.org/10.1080/14640749008401234

Epstein, M. L., & Emmanouil, T. A. (2017). Ensemble coding remains accurate under object and spatial visual working memory load. *Attention, Perception, & Psychophysics*, *79*, 2088-2097. https://doi.org/10.3758/s13414-017-1353-2

Esins, J., Bülthoff, I., & Schultz, J. (2011). *The role of featural and configural information for perceived similarity between faces* [Poster Presentation]. Annual Meeting of the Vision Sciences Society, Naples, Florida, USA. https://hdl.handle.net/11858/00-001M-0000-0013-BA94-8

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*(7), 589-601. https://doi.org/10.1068/i0669

Fahsing, I. A., Ask, K., & Granhag, P. A. (2004). The man behind the mask: Accuracy and predictors of eyewitness offender descriptions. *Journal of Applied Psychology*, *89*(4), 722-729. https://doi.org/10.1037/0021-9010.89.4.722

Finkbeiner, M., & Coltheart, M. (2009). Letter recognition: From perception to representation. *Cognitive Neuropsychology*, *26*(1), 1-6. https://doi.org/10.1080/02643290902905294

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389-1398. https://doi.org/10.1152/jn.00904.2010

Fiset, D., Blais, C., Arguin, M., Tadros, K., Éthier-Majcher, C., Bulb, D., & Gosselin, F. (2009). The spatio-temporal dynamics of visual letter recognition. *Cognitive Neuropsychology*, *26*(1), 23-35. https://doi.org/10.1080/02643290802421160

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law*, *19*(2), 151-164. https://doi.org/10.1037/a0030618

Flowe, H. D., & Ebbesen, E. B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. *Law and Human Behaviour*, *31*(1), 33-52. https://doi.org/10.1007/s10979-006-9045-9

Foroni, F., & Rothbart, M. (2011). Category boundaries and category labels: When does a category name influence the perceived similarity of category members? *Social Cognition*, *29*(5), 547-576. https://doi.org/10.1521/soco.2011.29.5.547

Furtak, M., Mudrik, L., & Bola, M. (2022). The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. *Cognition*, *221*, 1-7. https://doi.org/10.1016/j.cognition.2021.104983

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology*, *109*(2), 219-231. https://doi.org/10.1111/bjop.12260

Fysh, M. C., & Bindemann, M. (2023). Understanding face matching. *Quarterly Journal of Experimental Psychology*, *76*(4), 862-880. https://doi.org/10.1177/17470218221104476

Galfano, G., Sarlo, M., Sassi, F., Munafò, M., Fuentes, L. J., & Umiltà, C. (2011). Reorienting of spatial attention in gaze cuing is reflected in N2pc. *Social Neuroscience*, *6*(3), 257-269. https://doi.org/10.1080/17470919.2010.515722

Gawronski, B., & Quinn, K. A. (2013). Guilty by mere similarity: Assimilative effects of facial resemblance on automatic evaluation. *Journal of Experimental Social Psychology*, *49*(1), 120-125. https://doi.org/10.1016/j.jesp.2012.07.016

Gilad-Gutnick, S., Yovel, G., & Sinha, P. (2012). Recognizing degraded faces: The contribution of configural and featural cues. *Perception*, *41*(12), 1497-1511. https://doi.org/10.1068/p7064

Givens, G. H., Beveridge, J. R., Phillips, P. J., Draper, B., Lui, Y. M., & Bolme, D. (2013). Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics & Data Analysis*, *67*, 236-247. https://doi.org/10.1016/j.csda.2013.05.025

Goldenberg, A., Sweeny, T. D., Shpigel, E., & Gross, J. J. (2020). Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *Journal of Experimental Psychology: General*, *149*(3), 445-460. https://doi.org/10.1037/xge0000651

Griffiths, S., Rhodes, G., Jeffery, L., Palermo, R., & Neumann, M. F. (2018). The average facial expression of a crowd influences impressions of individual expressions. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(2), 311-319. https://doi.org/10.1037/xhp0000446

Guo, K., Meints, K., Hall, C., Hall, S., & Mills, D. (2009). Left gaze bias in humans, rhesus monkeys and domestic dogs. *Animal Cognition*, *12*(3), 409-418. https://doi.org/10.1007/s10071-008-0199-3

Guo, K., & Shaw, H. (2015). Face in profile view reduces perceived facial expression intensity: An eye-tracking study. *Acta Psychologica*, *155*, 19-28. https://doi.org/10.1016/j.actpsy.2014.12.001

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), 1-13. https://doi.org/10.1167/9.11.1

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751-R753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718-734. https://doi.org/10.1037/a0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825-1838. https://doi.org/10.3758/APP.72.7.1825

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5), 855-859. https://doi.org/10.3758/s13423-011-0125-6

Han, L., Leib, A. Y., Chen, Z., & Whitney, D. (2021). Holistic ensemble perception. *Attention, Perception, & Psychophysics*, *83*(3), 998-1013. https://doi.org/10.3758/s13414-020-02173-1

Hancock, P. J. B. (2021). Familiar faces as islands of expertise. *Cognition*, *214*, Article 104765. https://doi.org/10.1016/j.cognition.2021.104765

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330-337. https://doi.org/10.1016/S1364-6613(00)01519-9

Hansmann-Roth, S., Chetverikov, A., & Kristjánsson, Á. (2019). Representing color and orientation ensembles: Can observers learn multiple feature distributions? *Journal of Vision*, *19*(9), Article 2. https://doi.org/10.1167/19.9.2

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

Heyer, R., Chong, C., & Semmler, C. (2019). Facial image comparisons of morphed facial imagery. *Australian Journal of Forensic Sciences*, *51*, S5-S9. https://doi.org/10.1080/00450618.2019.1571106

Hills, P. J. (2018). Children process the self face using configural and featural encoding: Evidence from eye tracking. *Cognitive Development*, *48*, 82-93. https://doi.org/10.1016/j.cogdev.2018.07.002

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception*, *23*(1), 65-74. https://doi.org/10.1068/p230065

Honig, T., Shoham, A., & Yovel, G. (2022). Perceptual similarity modulates effects of learning from variability on face recognition. *Vision Research*, *201*, Article 108128. https://doi.org/10.1016/j.visres.2022.108128

Hoover, A. E. N., Démonet, J.-F., & Steeves, J. K. E. (2010). Superior voice recognition in a patient with acquired prosopagnosia and object agnosia. *Neuropsychologia*, *48*(13), 3725-3732. https://doi.org/10.1016/j.neuropsychologia.2010.09.008

Hsu, S.-M., & Lee, J.-S. (2016). Relative judgement in facial identity perception as revealed by sequential effects. *Attention, Perception, & Psychophysics*, *78*(1), 264-277. https://doi.org/10.3758/s13414-015-0979-1

Huang, T., Xiong, Z., & Zhang, Z. (2005). Face recognition applications. In S. Z. Li, & A. K. Jain (Eds.), *Handbook of face recognition.* Springer. https://doi.org/10.1007/0-387-27257-7_17

Hurlbert, A. (2007). Colour constancy. *Current Biology*, *17*(21), R906-R907. https://doi.org/10.1016/j.cub.2007.08.022

Im, H. Y., Chong, S. C., Sun, J., Steiner, T. G., Albohn, D. N., Adams, R. B., Jr., & Kveraga, K. (2017). Cross-cultural and hemispheric laterality effects on the ensemble coding of emotion in facial crowds. *Culture and Brain*, *5*(2), 125-152. https://doi.org/10.1007/s40167-017-0054-y

Jeckeln, G., Yavuzcan, S., Marquis, K. A., Sandipkumar Mehta, P., Yates, A. M., Phillips, P. J., & O'Toole, A. J. (2023). Human-machine comparison for cross-race face verification: Race bias at the upper limits of performance? arXiv:2305.16443v2.

Jeevan, G., Zacharias, G. C., Nair, M. S., & Rajan, J. (2022). An empirical study of the impact of masks on face recognition. *Pattern Recognition*, *122*, Article 108308. https://doi.org/10.1016/j.patcog.2021.108308

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science*, *319*(5862). 435. https://doi.org/10.1126/science.1149656

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

Jenkins, R., Lavie, N., & Driver, J. (2003). Ignoring famous faces: Category-specific dilution of distractor interference. *Perception & Psychophysics*, *65*(2), 298-309. https://doi.org/10.3758/BF03194801

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of Vision*, *18*(3), Article 17. https://doi.org/10.1167/18.3.17

Ji, L., & Hayward, W. G. (2021). Metacognition of average face perception. *Attention, Perception, & Psychophysics*, *83*(3), 1036-1048. https://doi.org/10.3758/s13414-020-02189-7

Jia, L., Cheng, M., Lu, J., Wu, Y., & Wang, J. (2023). Context consistency improves ensemble perception of facial expressions. *Psychonomic Bulletin & Review*, *30*(1), 280-290. https://doi.org/10.3758/s13423-022-02154-5

Jiang, Y., Shannon, R. W., Visueta, N., Bernat, E. M., Patrick, C. J., & He, S. (2009). Dynamics of processing invisible faces in the brain: Automatic neural encoding of facial expression information. *NeuroImage*, *44*(3), 1171-1177. https://doi.org/10.1016/j.neuroimage.2008.09.038

Jones, A. R., Carlson, C. A., Lockamyeir, R. F., Hemby, J. A., Carlson, M. A., & Wooton, A. R. (2020). “All I remember is the black eye”: A distinctive facial feature harms eyewitness identification. *Applied Cognitive Psychology*, *34*(6), 1379-1393. https://doi.org/10.1002/acp.3714

Jones, P. R., Moore, D. R., Shub, D. E., & Amitay, S. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1456-1470. https://doi.org/10.1037/xlm0000111

Kacin, M., Cha, O., & Gauthier, I. (2022). The relation between ensemble coding of length and orientation does not depend on spatial attention. *Vision*, *7*(1), 3. https://doi.org/10.3390/vision7010003

Kacin, M., Gauthier, I., & Cha, O. (2021). Ensemble coding of average length and average orientation are correlated. *Vision Research*, *187*, 94-101. https://doi.org/10.1016/j.visres.2021.04.010

Kanika., Singla, J., & Nikita. (2021). Comparing ROC curve based thresholding models in online transactions fraud detection system using deep learning. *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, *India*, 9-12. 10.1109/ICCCIS51004.2021.9397167

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*(8), 1163-1177. https://doi.org/10.3758/s13421-012-0226-0

Karaminis, T., Neil, L., Manning, C., Turi, M., Fiorentini, C., Burr, D., & Pellicano, E. (2017). Ensemble perception of emotions in autistic and typical children and adolescents. *Developmental Cognitive Neuroscience*, *24*, 51-62. https://doi.org/10.1016/j.dcn.2017.01.005

Keval, H. U., & Sasse, M. A. (2008). Can we ID from CCTV? Image quality in digital CCTV and face identification performance. *Proceedings of SPIE International Society for Optical Engineering*, *6982*, 69820K. https://doi.org/10.1117/12.774212

Khayat, N. & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), Article 23. https://doi.org/10.1167/18.9.23

Khvostov, V. A., Iakovlev, A. U., Wolfe, J. M., Utochkin, I. S. (2024). What is the basis of ensemble subset selection? *Attention, Perception, & Psychophysics*, *86*(3),776-798. https://doi.org/10.3758/s13414-024-02850-5

Kiss, M., Van Velzen, J., & Eimer, M. (2008). The N2pc component and its links to attention shifts and spatially selective visual processing. *Psychophysiology*, *45*(2), 240-249. https://doi.org/10.1111/j.1469-8986.2007.00611.x

Koca, Y., & Oriet, C. (2023). From pictures to the people in them: Averaging within-person variability leads to face familiarization. *Psychological Science*, *34*(2), 252-264. https://doi.org/10.1177/09567976221131520

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal component analysis. *Behavior Research Methods*, *49*(6), 2002-2011. https://doi.org/10.3758/s13428-016-0837-7

Kramer, R. S. S., Jenkins, R., Young, A. W., & Burton, A. M. (2017). Natural variability is essential to learning new faces. *Visual Cognition*, *25*(4-6), 470-476. https://doi.org/10.1080/13506285.2016.1242522

Kramer, R. S. S., Mireku, M. O., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, *4*, Article 28. https://doi.org/10.1186/s41235-019-0181-4

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*(6), 841-845. https://doi.org/10.1002/acp.3261

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, *15*(4), Article 1. https://doi.org/10.1167/15.4.1

Kuusikko, S., Haapsamo, H., Jansson-Verkasalo, E., Hurtig, T., Mattila, M.-L., Ebeling, H., Jussila, K., Bölte, S., & Moilanen, I. (2009). Emotion recognition in children and adolescents with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *39*(6), 938-945. https://doi.org/10.1007/s10803-009-0700-0

Lander, K., & Bruce, V. (2001). The role of motion in learning new faces. *Visual Cognition*, *10*(8), 897-912. https://doi.org/10.1080/13506280344000149

Large, M. E., & McMullen, P. A. (2006). Hierarchical attention in discriminating objects at different levels of specificity. *Perception & Psychophysics*, *68*(5), 845-860. https://doi.org/10.3758/BF03193706

Leder, H., Tinio, P. P. L., Fuchs, I. M., & Bohrn, I. (2010). When attractiveness demands longer looks: The effects of situation and gender. *Quarterly Journal of Experimental Psychology*, *63*(9), 1858-1871. https://doi.org/10.1080/17470211003605142

Lee, Y., Matsumiya, K., & Wilson, H. R. (2006). Size-invariant but viewpoint-dependent representation of faces. *Vision Research*, *46*(12), 1901-1910. https://doi.org/10.1016/j.visres.2005.12.008

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, *14*(8), Article 26. https://doi.org/10.1167/14.8.26

Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, *7*, Article 13186. https://doi.org/10.1038/ncomms13186

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia*, *50*(7), 1698-1707. https://doi.org/10.1016/j.neuropsychologia.2012.03.026

Li, H., Ji, L., Li, Q., & Chen, W. (2021). Individual faces were not discarded during extracting mean emotion representations. *Frontiers in Psychology*, *12*, Article 713212. https://doi.org/10.3389/fpsyg.2021.713212

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, *7*, Article 1332. https://doi.org/10.3389/fpsyg.2016.01332

Little, A. C., Hancock, P. J. B., DeBruine, L. M., & Jones, B. C. (2012). Adaptation to antifaces and the perception of correct famous identity in an average face. *Frontiers in Psychology*, *3*, Article 19. https://doi.org/10.3389/fpsyg.2012.00019

Liu, R., Ye, Q., Hao, S., Li, Y., Shen, L., & He, W. (2023). The relationship between ensemble coding and individual representation of crowd facial emotion. *Biological Psychology*, *180*, 1-11. https://doi.org/10.1016/j.biopsycho.2023.108593

Liu, Y., & Ji, L. (2024). Ensemble coding of multiple facial expressions is not affected by attentional load. *BMC Psychology*, *12*(1), Article 102. https://doi.org/10.1186/s40359-024-01598-9

Lobmaier, J. S., & Mast, F. W. (2007). Perception of novel faces: The parts have it! *Perception*, *36*(11), 1660-1673. https://doi.org/10.1068/p5642

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77-100. https://doi.org/10.1037/0096-1523.34.1.77

Longmore, C. A., Santos, I. M., Silva, C. F., Hall, A., Faloyin, D., & Little, E. (2017). Image dependency in the recognition of newly learnt faces. *Quarterly Journal of Experimental Psychology*, *70*(5), 863-873. https://doi.org/10.1080/17470218.2016.1236825

Lovén, J., Rehnman, J., Wiens, S., Lindholm, T., Peira, N., & Herlitz, A. (2012). Who are you looking at? The influence of face gender on visual attention and memory for own- and other-race faces. *Memory*, *20*(4), 321-331. https://doi.org/10.1080/09658211.2012.658064

Lucas, C. A., Neil, B., & Palmer, M. A. (2021). Eyewitness identification: The complex issue of suspect-filler similarity. *Psychology, Public Policy, and Law*, *27*(2), 151-169. https://doi.org/10.1037/law0000243

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS ONE*, *5*(7), e11452. https://doi.org/10.1371/journal.pone.0011452

MacLin, O. H., MacLin, M. K., & Malpass, R. S. (2001). Race, arousal, attention, exposure and delay: An examination of factors moderating face recognition. *Psychology, Public Policy, and Law*, *7*(1), 134-152. https://doi.org/10.1037/1076-8971.7.1.134

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*(9-10), 447-458. https://doi.org/10.1080/13506285.2017.1301614

Mäntylä, T., & Sundström, A. (2004). Changing scenes: Memory for naturalistic events following change blindness. *Memory*, *12*(6), 696-706. https://doi.org/10.1080/09658210344000585

Matthews, C. M., Davis, E. E., & Mondloch, C. J. (2018). Getting to know you: The development of mechanisms underlying face learning. *Journal of Experimental Child Psychology*, *167*, 295-313. https://doi.org/10.1016/j.jecp.2017.10.012

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4), Article 6. https://doi.org/10.1167/15.4.6

Megreya, A. M., & Bindemann, M. (2012). Identification accuracy for single- and double-perpetrator crimes: Does accomplice gender matter? *British Journal of Psychology*, *103*(4), 439-453. https://doi.org/10.1111/j.2044-8295.2011.02084.x

Megreya, A. M., & Burton, A. M. (2006a). Recognising faces seen alone or with others: When two heads are worse than one. *Applied Cognitive Psychology*, *20*(7), 957-972. https://doi.org/10.1002/acp.1243

Megreya, A. M., & Burton, A. M. (2006b). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. https://doi.org/10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364-372. https://doi.org/10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700-706. https://doi.org/10.1002/acp.2965

Meinhardt, G., Meinhardt-Injac, B., Persike, M. (2014). The complete design in the composite face paradigm: Role of response bias, target certainty, and feedback. *Frontiers in Human Neuroscience*, *8*, Article 885. https://doi.org/10.3389/fnhum.2014.00885

Meinhardt-Injac, B., Boutet, I., Persike, M., Meinhardt, G., & Imhof, M. (2017). From development to aging: Holistic face perception in children, younger and older adults. *Cognition*, *158*, 134-146. https://doi.org/10.1016/j.cognition.2016.10.020

Menon, N., Kemp, R. I., & White, D. (2018). More than a sum of parts: Robust face recognition by integrating variation. *Royal Society Open Science*, *5*(5), 172381. https://doi.org/10.1098/rsos.172381

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception*, *44*(11), 1332-1341. https://doi.org/10.1177/0301006615599902

Mestry, N., Menneer, T., Cave, K. R., Godwin, H. J., & Donnelly, N. (2017). Dual-target cost in visual search for multiple unfamiliar faces. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1504-1519. https://doi.org/10.1037/xhp0000388

Neal, D. (2023). *A re-inspection of ePassport gates*. Independent Chief Inspector of Borders and Immigration. https://assets.publishing.service.gov.uk/media/65e075ed2f2b3b001c7cd769/A_re-inspection_of_ePassport_gates_May_2023.pdf

Neumann, M. F., De Bonis, F., Rhodes, G., & Palermo, R. (2015). The role of similarity in coding ensemble identity of face groups. *Journal of Vision*, *15*(12), 705. https://doi.org/10.1167/15.12.705

Neumann, M. F., Ng. R., Rhodes, G., & Palermo, R. (2018). Ensemble coding of face identity is not independent of the coding of individual identity. *Quarterly Journal of Experimental Psychology*, *71*(6), 1357-1366. https://doi.org/10.1080/17470218.2017.1318409

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, *128*(1), 56-63. https://doi.org/10.1016/j.cognition.2013.03.006

Nightingale, S. J., Agarwal, S., & Farid, H. (2021). Perceptual and computational detection of face morphing. *Journal of Vision*, *21*(3), Article 4. https://doi.org/10.1167/jov.21.3.4

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences*, *9*(10), 467-473. https://doi.org/10.1016/j.tics.2005.08.004

Norman, L. J., & Tokarev, A. (2014). Spatial attention does not modulate holistic face processing, even when multiple faces are present. *Perception*, *43*(12), 1341-1352. https://doi.org/10.1068/p7848

Nortje, A., Tredoux, C. G., & Vredeveldt, A. (2017). How many faces can we remember? Why this matters when assessing eyewitnesses. In M. Bindemann & A. Megreya (Eds.), *Face processing: Systems, disorders and cultural disorders*. NOVA Science Publishers.

Oderkerk, C. A. T., & Beier, S. (2024). Script-style degrees: Letter recognition in regular versus special fonts. *Information Design Journal*, *29*(1), 25-35. https://doi.org/10.1075/idj.22021.ode

Oh, B.-I., Kim, Y.-J., & Kang, M.-S. (2019). Ensemble representations reveal distinct neural coding of visual working memory. *Nature Communications*, *10*(1), 5665. https://doi.org/10.1038/s41467-019-13592-6

Oliva, A., & Schyns, P. G. (2000). Diagnostic color mediates scene recognition. *Cognitive Psychology*, *41*(2), 176-210. https://doi.org/10.1006/cogp.1999.0728

O'Toole, A. J., An, X., Dunlop, J., Natu, V., & Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception*, *9*(4), 1-13. https://doi.org/10.1145/2355598.235559

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(9), 1642-1646. https://doi.org/10.1109/tpami.2007.1107

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*, *51*(19), 2145-2155. https://doi.org/10.1016/j.visres.2011.08.009

Palermo, R., & Rhodes, G. (2002). The influence of divided attention on holistic face perception. *Cognition*, *82*(3), 225-257. https://doi.org/10.1016/S0010-0277(01)00160-3

Park, S., Kim, M.-S., & Chun, M. M. (2007). Concurrent working memory load can facilitate selective attention: Evidence for specialized load. *Journal of*

Experimental Psychology: Human Perception and Performance, *33*(5), 1062-1075. https://doi.org/10.1037/0096-1523.33.5.1062

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Peng, S., Kuang, B., & Hu, P. (2019). Memory of ensemble representation was independent of attention. *Frontiers in Psychology*, *10*, Article 228. https://doi.org/10.3389/fpsyg.2019.00228

Peng, S., Liu, C. H., & Hu, P. (2021). Effects of subjective similarity and culture on ensemble perception of faces. *Attention, Perception, & Psychophysics*, *83*(3), 1070-1079. https://doi.org/10.3758/s13414-020-02133-9

Peng, S., Liu, C. H., Liu, W., & Yang, Z. (2022). Emotion matters: Face ensemble perception is affected by emotional states. *Psychonomic Bulletin & Review*, *29*(1), 116-122. https://doi.org/10.3758/s13423-021-01987-w

Peng, S., Zhang, L., Xu, R., Liu, C. H., Chen, W., & Hu, P. (2019). Self-construal priming modulates ensemble perception of multiple-face identities. *Frontiers in Psychology*, *10*, Article 1096. https://doi.org/10.3389/fpsyg.2019.01096

Phillips, P. J., & O’Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, *31*(1), 74-85. https://doi.org/10.1016/j.imavis.2013.12.002

Phillips, P. J., Yates, A. N., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O’Toole, A., (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms.

Proceedings of the National Academy of Sciences of the United States of America, *115*(24), 6171-6176. https://doi.org/10.1073/pnas.1721355115

Platek, S. M., Critton, S. R., Burch, R. L., Frederick, D. A., Myers, T. E., & Gallup Jr., G. G. (2003). How much paternal resemblance is enough? Sex differences in hypothetical investment decisions but not in the detection of resemblance. *Evolution and Human Behavior*, *24*(2), 81-87. https://doi.org/10.1016/S1090-5138(02)00117-4

Polk, T. A., Lacey, H. P., Nelson, J. K., Demiralp, E., Newman, L. I., Krauss, D. A., Raheja, A., & Farah, M. J. (2009). The development of abstract letter representations for reading: Evidence for the role of context. *Cognitive Neuropsychology*, *26*(1), 70-90. https://doi.org/10.1080/02643290802618757

Popova, T., & Wiese H. (2022). The time it takes to truly know someone: Neurophysiological correlates of face and identity learning during the first two years. *Biological Psychology*, *170*, 108312. https://doi.org/10.1016/j.biopsycho.2022.108312

Popova, T., & Wiese, H. (2023). Developing familiarity during the first eight months of knowing a person: A longitudinal EEG study on face and identity learning. *Cortex*, *165*, 26-37. https://doi.org/10.1016/j.cortex.2023.04.008

Portch, E., Wignall, L., & Bate, S. (2023). Why can some people with developmental prosopagnosia recognise some familiar faces? Insights from subjective experience. *PeerJ*, *11*, e15497. https://doi.org/10.7717/peerj.15497

Qarooni, R., Prunty, J., Bindemann, M., & Jenkins, R. (2022). Capacity limits in face detection. *Cognition*, *228*, 1-8. https://doi.org/10.1016/j.cognition.2022.105227

Rajendran, S., Maule, J., Franklin, A., & Webster, M. A. (2021). Ensemble coding of colour and luminance contrast. *Attention, Perception, & Psychophysics*, *83*, 911-924. https://doi.org/10.3758/s13414-020-02136-6

Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Christensen, A. (1990). The unconscious transference effect: Are innocent bystanders ever misidentified? *Applied Cognitive Psychology*, *4*(1), 3-31. https://doi.org/10.1002/acp.2350040103

Reynolds, J. K., & Pezdek, K. (1992). Face recognition memory: The effects of exposure duration and encoding instruction. *Applied Cognitive Psychology*, *6*(4), 279-292. https://doi.org/10.1002/acp.2350060402

Rhodes, G., Neumann, M. F., Ewing, L., Bank, S., Read, A., Engfors, L. M., Emiechel, R., & Palermo, R. (2018). Ensemble coding of faces occurs in children and develops dissociably from coding of individual faces. *Developmental Science*, *21*(2), e12540. https://doi.org/10.1111/desc.12540

Rhodes, G., Neumann, M. F., Ewing, L., & Palermo, R. (2015). Reduced set averaging of face identity in children and adolescents with autism. *Quarterly Journal of Experimental Psychology*, *67*(7), 1391-1403. https://doi.org/10.1080/17470218.2014.981554

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464-471. https://doi.org/10.1177/0956797611401753

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision Research*, *49*(23), 2856-2861. https://doi.org/10.1016/j.visres.2009.08.025

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897-905. https://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Flack, T. R., & Maréchal, L. (2022). Unfamiliar faces might as well be another species: Evidence from a face matching task with human and monkey faces. *Visual Cognition*, *30*(10), 680-685. https://doi.org/10.1080/13506285.2023.2184894

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (2017). Face averages and multiple images in a live matching task. *British Journal of Psychology*, *111*(1), 92-102. https://doi.org/10.1111/bjop.12388

Ritchie, K. L., White, D., Kramer, R. S. S., Noyes, E., Jenkins, R., & Burton, A. M. (2018). Enhancing CCTV: Averages improve face identification from poor-quality images. *Applied Cognitive Psychology*, *32*(6), 671-680. https://doi.org/10.1002/acp.3449

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS ONE*, *12*(3), Article e0173319. https://doi.org/10.1371/journal.pone.0173319

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE*, *11*(2), Article e0150036. https://doi.org/10.1371/journal.pone.0150036

Robitaille, N., & Jolicoeur, P. (2006). Fundamental properties of the N2pc as an index of spatial attention: Effects of masking. *Canadian Journal of Experimental Psychology*, *60*(2), 101-111. https://doi.org/10.1037/cjep2006011

Robson, M. K., Palermo, R., Jeffery, L., & Neumann, M. F. (2018). Ensemble coding of face identity is present but weaker in congenital prosopagnosia. *Neuropsychologia*, *111*, 377-386. https://doi.org/10.1016/j.neuropsychologia.2018.02.019

Rose, C., & Beck, V. (2016). Eyewitness accounts: False facts, false memories, and false identification. *Journal of Crime and Justice*, *39*(2), 243-263. https://doi.org/10.1080/0735648X.2014.940999

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*(1), 107-113. https://doi.org/10.1038/nn1370

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get the “gist” of real-world natural scenes? *Visual Cognition*, *12*(6), 852-877. https://doi.org/10.1080/13506280444000553

Rump, K. M., Giovannelli, J. L., Minshew, N. J., & Strauss, M. S. (2009). The development of emotion recognition in individuals with autism. *Child Development*, *80*(5), 1434-1447. https://doi.org/10.1111/j.1467-8624.2009.01343.x

Russ, A. J., Sauerland, M., Lee, C. E., & Bindemann, M. (2018). Individual differences in eyewitness accuracy across multiple lineups of faces. *Cognitive Research: Principles and Implications*, *3*, Article 30, https://doi.org/10.1186/s41235-018-0121-8

Sama, M. A., Nestor, A., & Cant, J. S. (2019). Independence of viewpoint and identity in face ensemble processing. *Journal of Vision*, *19*(5), Article 2. https://doi.org/10.1167/19.5.2

Sandford, A., & Ritchie, K. L. (2021). Unfamiliar face matching, within-person variability, and multiple-image arrays. *Visual Cognition*, *29*(3), 143-157. https://doi.org/10.1080/13506285.2021.1883170

Sanocki, T. (1993). Time course of object identification: Evidence for a global-to-local contingency. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 878-898. https://doi.org/10.1037/0096-1523.19.4.878

Schweinberger, S. R., Pickering, E. C., Burton, A. M., & Kaufmann, J. M. (2002). Human brain potential correlates of repetition priming in face and name recognition. *Neuropsychologia*, *40*(12), 2057-2073. https://doi.org/10.1016/S0028-3932(02)00050-7

Sexton, L., Mileva, M., Hole, G., Strathie, A., & Laurence, S. (2023). Recognizing newly learned faces across changes in age. *Visual Cognition*, *31*(8), 617-632. https://doi.org/10.1080/13506285.2024.2315813

Shelke, V., Mehta, G., Gomase, P., & Bangera, T. (2021). Searchious: Locating missing people using an optimised face recognition algorithm. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1550-1555. 10.1109/ICCMC51019.2021.9418450

Song, G. (2013). Comparison of visual similarity judgement on computer screens and on paper. *Theoretical Issues in Ergonomics Science*, *14*(2), 126-137. https://doi.org/10.1080/1464536X.2011.584583

Stabile, V. J., Baker, K. A., & Mondloch, C. J. (2024). Criterion shifting in an unfamiliar face-matching task: Effects of base rates, payoffs, and perceptual discriminability. *Journal of Applied Research in Memory and Cognition*, *13*(4), 569-581. https://doi.org/10.1037/mac0000157

Stacey, P. C., Walker, S., & Underwood, J. D. M. (2005). Face processing and familiarity: Evidence from eye-movement data. *British Journal of Psychology*, *96*(4), 407-422. https://doi.org/10.1348/000712605X47422

Stantić, M., Brown, K., Ichijo, E., Pounder, Z., Catmur, C., & Bird, G. (2023). Independent measurement of face perception, face matching, and face memory reveals impairments in face perception and memory, but not matching, in autism. *Psychonomic Bulletin & Review*, *30*(6), 2240-2249. https://doi.org/10.3758/s13423-023-02304-3

Stantić, M., Pounder, Z., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2022). Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *Cortex*, *157*, 266-273. https://doi.org/10.1016/j.cortex.2022.09.012

Stephan, B. C. M., & Caine, D. (2007). What is in a view? The role of featural information in the recognition of unfamiliar faces across viewpoint transformation. *Perception*, *36*(2), 189-198. https://doi.org/10.1068/p5627

Stephan, C. N., & Arthur, R. S. (2006). Assessing facial approximation accuracy: How do resemblance ratings of disparate faces compare to recognition tests? *Forensic Science International*, *159*, S159-S163. https://doi.org/10.1016/j.forsciint.2006.02.026

Stevenage, S. V., Hale, S., Morgan, Y., & Neil, G. J. (2014). Recognition by association: Within- and cross-modality associative priming with faces and voices. *British Journal of Psychology*, *105*(1), 1-16. https://doi.org/10.1111/bjop.12011

Sun, J., & Chong, S. C. (2020). Power of averaging: Noise reduction by ensemble coding of multiple faces. *Journal of Experimental Psychology: General*, *149*(3), 550-563. https://doi.org/10.1037/xge0000667

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behaviour: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 329-337. https://doi.org/10.1037/a0028712

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4-5-year-old children. *Developmental Science*, *18*(4), 556-568. https://doi.org/10.1111/desc.12239

Swystun, A. G., & Logan, A. J. (2019). Quantifying the effect of viewpoint changes on sensitivity to face identity. *Vision Research*, *165*, 1-12. https://doi.org/10.1016/j.visres.2019.09.006

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. *Cognition*, *93*(1), B1-B9. https://doi.org/10.1016/j.cognition.2003.09.011

Thoma, V., & Lavie, N. (2013). Perceptual load effects on processing distractor faces indicate face-specific capacity limits. *Visual Cognition*, *21*(8), 1053-1076. https://doi.org/10.1080/13506285.2013.853717

Thompson, W. B., & Johnson, J. (2008). Biased lineup instructions and face identification from video images. *The Journal of General Psychology*, *135*(1), 23-36. https://doi.org/10.3200/GENP.135.1.23-36

Thornton, I. M., Srismith, D., Oxner, M., & Hayward, W. G. (2019). Other-race faces are given more weight than own-race faces when assessing the composition of crowds. *Vision Research*, *157*, 159-168. https://doi.org/10.1016/j.visres.2018.02.008

Tomita, A., Yamamoto, S., Matsushita, S., & Morikawa, K. (2014). Resemblance to familiar faces is exaggerated in memory. *Japanese Psychological Research*, *56*(1), 24-32. https://doi.org/10.1111/jpr.12032

Trinh, A., Dunn, J. D., & White, D. (2022). Verifying unfamiliar identities: Effects of processing name and face information in the same identity-matching task. *Cognitive Research: Principles and Implications*, *7*, Article 92. https://doi.org/10.1186/s41235-022-00441-2

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352. https://doi.org/10.1037/0033-295X.84.4.327

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision*, *15*(4), Article 8. https://doi.org/10.1167/15.4.8

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journel of Experimental Psychology*, *69*(10), 1997-2019. https://doi.org/10.1080/17470218.2014.990392

Van Belle, G., De Graef, P., Verfaillie, K., Busigny, T., & Rossion, B. (2010). Whole not hole: Expert face recognition requires holistic perception. *Neuropsychologia*, *48*(9), 2620-2629. https://doi.org/10.1016/j.neuropsychologia.2010.04.034

Vladeanu, M., Lewis, M., & Ellis, H. (2006). Associative priming in faces: Semantic relatedness or simple co-occurrence? *Memory & Cognition*, *34*(5), 1091-1101. https://doi.org/10.3758/BF03193255

Wammes, J. D, & Fernandes, M. A. (2016). Interfering with memory for faces: The cost of doing two things at once. *Memory*, *24*(2), 184-203. https://doi.org/10.1080/09658211.2014.998240

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*(5), 553-571. https://doi.org/10.1037/0003-066X.48.5.553

White, M., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE*, *9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

White, M., Wayne, T., & Varela, V. P. L. (2022). Partitioning natural face image variability emphasises within-identity over between-identity representation for understanding accurate recognition. *Cognition*, *219*, 1-11. https://doi.org/10.1016/j.cognition.2021.104966

Wieser, M. J., Hambach, A., & Weymar, M. (2018). Neurophysiological correlates of attentional bias for emotional faces in socially anxious individuals – Evidence from a visual search task and N2pc. *Biological Psychology*, *132*, 192-201. https://doi.org/10.1016/j.biopsycho.2018.01.004

Wiesmann, S. L., & Võ, M. L.-H. (2022). What makes a scene? Fast scene categorization as a function of global scene information at different resolutions. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 871-888. https://doi.org/10.1037/xhp0001020

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research*, *24*(1), 55-62. https://doi.org/10.1016/0042-6989(84)90144-5

Wolfe, B. A., Kosovicheva, A. A., Leib, A. Y., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision*, *15*(4), Article 11. https://doi.org/10.1167/15.4.11

Wu, E. X. W., Laeng, B., & Magnussen, S. (2012). Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition. *Social Neuroscience*, *7*(2), 202-216. https://doi.org/10.1080/17470919.2011.596946

Yang, J.-W., Yoon, K. L., Chong, S. C., & Oh, K. J. (2013). Accurate but pathological: Social anxiety and ensemble coding of emotion. *Cognitive Therapy and Research*, *37*(3), 572-578. https://doi.org/10.1007/s10608-012-9500-5

Yang, Z., Wu, J., Liu, S., Zhao, L., Fan, C., & He, W. (2024). Ensemble coding of crowd with cross-category facial expressions. *Behavioural Sciences*, *14*(6), 508. https://doi.org/10.3390/bs14060508

Ying, H. (2022). Attention modulates the ensemble coding of facial expressions. *Perception*, *51*(4), 276-285. https://doi.org/10.1177/03010066221079686

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, Article 104135. https://doi.org/10.1016/j.cognition.2019.104135

Zhou, G., Cheng, Z., Yue, Z., Tredoux, C., He, J., & Wang, L. (2015). Own-race faces capture attention faster than other-race faces: Evidence from response time and the N2pc. *PLoS ONE*, *10*(6), Article e0127709. https://doi.org/10.1371/journal.pone.0127709

Zhou, Y., Liu, D., & Huang, T. (2018). Survey of face detection on low-quality images. In *2018 13th IEEE International Conference on Automatic Face*

Gesture Recognition (pp. 769-773).

https://doi.org/10.48550/arXiv.1804.07362