Investigating face perception in humans and DCNNs

Author:
Perrone de Lima Varela, Victor

Publication Date:
2023

DOI:
https://doi.org/10.26190/unsworks/25206

License:
https://creativecommons.org/licenses/by/4.0/
Link to license to see what you are allowed to do with this resource.

Downloaded from http://hdl.handle.net/1959.4/101499 in https://unsworks.unsw.edu.au on 2024-04-27

Investigating face perception in humans and DCNNs

Víctor Perrone de Lima Varela

A thesis in fulfilment of the requirements for the degree

of Doctor of Philosophy

School of Psychology

Faculty of Science

August 2023

Declarations

ORIGINALITY STATEMENT

☑ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

☑ I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Publications Statement

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **some of the work described in their thesis has been published and has been documented in the relevant Chapters with acknowledgement**.

A short statement on where this work appears in the thesis and how this work is acknowledged within chapter/s:

> Chapters 4 and 5 contain experimental work that was either accepted or published in academic journals. These publications are acknowledged on page 10 of the thesis document under the heading "Publications from this thesis".

Candidate's Declaration

I declare that I have complied with the Thesis Examination Procedure.

Table of Contents

Publications from this thesis 11

Acknowledgements 12

Abstract 13

Chapter 1 - General Introduction 14

Human performance in face identity processing tasks 16

How face images affect unfamiliar face matching performance in applied settings 17

Individual differences in identity processing 19

Improving human performance 21

Technological alternatives 23

Human-AI hybrid approaches 26

Using DCNNs as a model of human processing: comparing man vs machine 27

Using Eye-tracking to understand processing differences underlying face identity processing in humans 29

Lab-based eye tracking research and ecological validity reflecting human face processing 32

Thesis aims 33

Chapter 2 - Comparing human and machine performance when matching images of different quality 35

Introduction 35

Experiment 1 40

Method 41

Participants 41

Deep Convolutional Neural Networks (DCNNs) 41
Stimuli 43
Procedure 47
Results....49
Analysis of accuracy....49
Analysis of similarity ratings 52
Analysis of feature similarity space in DCNNs....54
Discussion 55
Conclusion....60
Chapter 3 - Optimising human and AI teaming by being sensitive to individual differences .62
Introduction....62
Computational study 1 66
Method....67
Participants....67
Stimuli 67
Procedure 67
Results....70
Human-DCNN fusion improvements correlate with human performance 70
Human-DCNN fusion is improved when human performance is within 10% of DCNNs' performance....72
Disagreements between humans and DCNN improve fusion performance 75
Discussion 78
Conclusion....82

Chapter 4 - Face information use in humans, super recognisers, and DCNNs 83
Introduction.....83
Individual differences and face information use.....83
Eye-tracking studies of information use.....85
Our approach.....87
Chapter Objectives.....89
Experiment 2.....90
Methods.....92
Participants.....92
Stimuli.....92
Procedure.....92
Eye movement classification.....93
Analysis of fixation patterns.....93
Principal components analysis (PCA) of Heatmaps.....93
Maximum uncertainty Linear Discriminant Analysis (MLDA) of Heatmaps.....94
Regions-Of-Interest.....95
Gini coefficient as a measure of gaze dispersal in heatmaps.....97
Results.....97
Overall Accuracy.....97
Comparison of heatmap and ROI approaches.....98
Inter-individual difference analysis of fixation patterns.....99
Intra-individual difference analysis of fixation patterns.....102
Individual difference analysis of heatmaps using MLDA.....104

Individual difference analysis of visual exploration (Gini coefficient) 107
Discussion 109
Experiments 3A & 3B 111
Experiment 3A 111
Methods 112
Participants 112
Stimuli 112
Apparatus and eye movement classification 113
Procedure 113
Fixation Analysis: Heatmaps 113
Exploration analysis: Gini coefficient 114
Results 114
Overall accuracy 114
Principal Component Analysis 115
Principal Component Analysis (Natural Viewing condition only) 118
Intra-individual difference analysis of fixation patterns (Natural Viewing condition only) 120
Gaze dispersal Analysis (Natural Viewing Only) 122
Discussion 123
Experiment 3B 125
Methods 126
Participants 126
Stimuli 126

Apparatus and eye movement classification 126
Results....126
Overall Accuracy 126
Principal Components Analysis....127
Principal Components Analysis (Natural Viewing only)....128
Intra-individual difference analysis of fixation patterns (Natural Viewing only)....129
Gaze dispersal Analysis (Natural Viewing only) 132
Discussion....132
Experiment 4 133
Methods....135
Deep Convolutional Neural Networks (DCNNs) 135
Stimuli 136
Analysis....139
Results....140
Overall Accuracy 140
Information available 141
Discussion....143
Chapter discussion....144
Chapter 5 - Using DCNNs to examine human face perception in the wild....149
Introduction....149
Experiment 5 151
Methods....152
Participants....152

Apparatus and eye movement classification 152
Procedure 152
Eye gaze data processing 154
Comparing automatic versus manual coding 155
Navigation task 155
Face diet 156
Face-to-Face interaction task 156
Results 156
Faces of passersby do not capture attention in live natural settings 156
Influence of social attention on ‘face diet’ 159
Individual differences in naturalistic social attention 160
Fixation patterns during face-to-face interaction associated with face recognition ability 162
Discussion 166
Chapter 6 - General Discussion 169
Summary of research aims and findings 169
Main Findings 172
Face processing tasks are difficult for humans and DCNNs 172
Humans and DCNNs possess different strategies for processing identity information 172
Humans and DCNNs used in conjunction improve the quality of face verification decisions 173
Information sampling provides routes for expertise in face-processing tasks 174
Human-guided information sampling benefits DCNNs for facial identity information 175

Social attention in the wild conflicts with screen-based research....176
Practical Implications....178
Methodological Implications....180
Theoretical implications....181
Conclusion....182
References....183
APPENDIX A....211
Method for pilot study to choose frequency bands in Chapter 2....211
Descriptive Table....211
ANOVA Table....217
Similarity Scores Distributions....221
t-SNE visualisation....223
APPENDIX B....226
Results for Gaussian Blur....226
APPENDIX C....235
Experiment 2 - Results for trial-level data....235
Experiment 3A....241
Aperture size determination....241
PCA – Learning and Recognition phases....242
Gaze dispersal Analysis (all apertures)....248
PCA – Learning and Recognition (Only NV)....249
Experiment 3B....252
PCA – Learning and Recognition....252

Gaze Dispersal Analysis (all apertures) 255
PCA – Learning and Recognition (Only NV) 255
Experiment 4 257
AUC Analysis 257
Information available Analysis 257
APPENDIX D 259
Individual visualisation of participants' body maps during the navigation task 259
Extended ANOVA analysis for 'Faces of passersby do not capture attention in live natural settings' 262
Extended analysis for 'Individual differences in naturalistic social attention' (analysis of residuals) 262
Individual visualisation of participants' facial maps during the face-to-face interaction task 263
Comparing automatic versus manual coding 267
References 268

Publications from this thesis

Chapter 4: Experiments 3A – 3B

Dunn, J. D., Varela, V. P. L., Nicholls, V. I., Papinutto, M., White, D., & Miellet, S. (2022). Face-Information Sampling in Super-Recognizers. *Psychological Science*, *33*(9), 1615-1630.

I contributed to all aspects of this paper. But my unique contribution, and the focus of the reporting of Experiments 3A and 3B that is included in this thesis, was to apply and extend analysis approaches that I developed in my masters work to explore individual differences in face information use in human participants. For the purpose of Chapter 4, I have written up this work in the context of my independent line of the investigation reported in this chapter.

Chapter 5: Experiment 5

Varela, V. P., Towler, A., Kemp, R. I., & White, D. (2023). Looking at faces in the wild. *Scientific Reports*, *13*(1), 783.

Although this was a collaboration with my supervision team, as the lead author on this paper, I contributed heavily to project design, and I had sole responsibility for conducting the study, the technical work necessary for the methodological advance reported here and the analysis. I produced the first draft of the manuscript and incorporated critical revisions made by the other authors. As such, in this thesis, the manuscript is presented in the form that it was submitted to the journal prior to revision and resubmission decision.

Acknowledgements

Writing this section feels very nostalgic and emotional. At the beginning of my four-year PhD journey, I was just a silly boy from the depths of a Brazilian forest. Now, it ends with a grown man open to the whole world, who I dare say is more mature, responsible, and resilient. So, I think it is fair to start this section by thanking the person I was from the past for having the guts to accept and embark on such an insane journey to the unknown. Thanks, mate! It turned out to be great.

Ana Maria Varela and Waldir Varela, NONE of this would have been possible without your daily support. Thanks for being awesome parents!

Camila Varela, no words can describe how lucky I am to have a person like you in my life. Thanks for giving me the emotional strength to pursue this challenge and constantly reminding me that everything is possible as long as we have each other. I see you, and I love you.

David White, thanks for making this journey possible by accepting me as your student. I will forever be grateful for all the fantastic things you allowed me to see and experience.

Richard Kemp and Penelope Earp, thanks for opening your house door and making me feel at home on the other side of the planet. I will forever be grateful for your kindness and trust.

My mentors Alice Towler, Carlos Thomaz, David White, James Dunn, Richard Kemp, and Sebastien Miellet. Thanks for believing in me, accepting me, and patiently showing me how to make our art reach its finest. You all inspire me to keep becoming a better version of myself.

My lab colleagues Anita Trinh, Bojana Popovic, Daniel Guilbert, Rebecca Tyler, Mariam Younan, Monique Piggot, Stephanie Summersby, and countless others. Thanks for being good friends and for cheering me up when I most needed it. You are all incredible human beings!

My flatmates Aadarsh Subramani and Melody Durupt, thanks for all the gaming nights and for tolerating me living with you for long years. I promise that as soon as I am done with this thing, I will move out (or maybe not, haha). Love you, guys!

My lifelong friends who somehow helped me through this journey, thank you so much! It is boring to say thanks without mentioning people. And so, a shout-out to Rodrigo Carteiro, Paula Nunes, Carolina Carteiro, Ivan Cerqueira, Leticia Salatiel, Juliana Bonardi, Isabela Salina, and Chicco Mattos and crew.

Sorry if I forgot to put your name here. It would take multiple pages to do so. As a solution: Thanks to _____ for being part of this journey!

Abstract

This thesis aims to compare strengths and weaknesses of AI and humans performing face identification tasks, and to use recent advances in machine-learning to develop new techniques for understanding face identity processing. By better understanding underlying processing differences between Deep Convolutional Neural Networks (DCNNs) and humans, it can help improve the ways in which AI technology is used to support human decision-making and deepen understanding of face identity processing in humans and DCNNs. In Chapter 2, I test how the accuracy of humans and DCNNs is affected by image quality and find that humans and DCNNs are affected differently. This has important applied implications, for example, when identifying faces from poor-quality imagery in police investigations, and also points to different processing strategies used by humans and DCNNs. Given these diverging processing strategies, in Chapter 3, I investigate the potential for human and DCNN decisions to be combined in face identification decisions. I find a large overall benefit of 'fusing' algorithm and human face identity judgments, and that this depends on the idiosyncratic accuracy and response patterns of the particular DCNNs and humans in question. This points to new optimal ways that individual humans and DCNNs can be aggregated to improve the accuracy of face identity decisions in applied settings.

Building on my background in computer vision, in Chapters 4 and 5, I then aim to better understand face information sampling by humans using a novel combination of eye-tracking and machine-learning approaches. In chapter 4, I develop exploratory methods for studying individual differences in face information sampling strategies. This reveals differences in the way that 'super-recognisers' sample face information compared to typical viewers. I then use DCNNs to assess the computational value of the face information sampled by these two groups of human observers, finding that sampling by 'super-recognisers' contains more computationally valuable face identity information. In Chapter 5, I develop a novel approach to measuring fixations to people in unconstrained natural settings by combining wearable eye-tracking technology with face and body detection algorithms. Together, these new approaches provide novel insight into individual differences in face information sampling, both when looking at faces in lab-based tasks performed on computer monitors and when looking at faces 'in the wild'.

Chapter 1 - General Introduction

Understanding the signals emitted by faces is a vital task for humans. Extracting and interpreting cues relating to, for example, biological sex, emotional state, gaze direction, and identity is critical to recognising and interacting successfully with one another. Humans have therefore developed a functioning system tuned to understanding the visual cues emitted by the faces around us.

This tuned system has led to the theoretical position that people are 'experts' in processing the signals emitted by faces (e.g. Carey, 1992). However, a significant exception to this is when people attempt to match the identity of unfamiliar face images. For processing face identity information, our expertise is reserved only for familiar people, as unfamiliar face matching is very challenging (see Young & Burton, 2018). And so, this lack of expertise in processing unfamiliar faces is problematic given the importance of this task in applied settings, for example, court proceedings and applied security settings. Numerous studies now show large proportions of errors, including in professional populations that perform the task in daily work (see White, Towler, & Kemp, 2021).

Recent technological advancements mean that it is increasingly common to see these settings implement automatic systems for face recognition based on Artificial intelligence (AI) algorithms to replace - or support – human decisions. The new generation of Deep Convolutional Neural Networks (DCNNs) - effortlessly operating today - can now match pairs of unfamiliar face images by identity with similar accuracy to the best humans (e.g. Phillips et al., 2018). However, despite being powerful tools, such algorithms are still subject to errors in unfamiliar face matching, raising similar concerns compared to humans regarding their use in security settings.

One of this thesis aims is to compare the accuracy of humans and DCNNs and the strengths and weaknesses of them performing similar tasks. By better understanding the underlying processing differences between humans and DCNNs processing faces, we argue that it can help improve how AI technology is used to support human decision-making, providing practical gains regarding the accuracy - and fairness - of face identification systems using these tools. In addition to this practical outcome, this aim can also help deepen the understanding of face identity processing in humans because DCNNs have recently received significant attention as potential candidates to model face identity

processing in humans. And so, comparing the underlying similarities of human and DCNN processing also tests how well the current generation of DCNN models human perceptual and cognitive processing.

The second aim of this thesis is to use my engineering expertise for methodological gains. Given my background in Electrical and Computer Engineering, I explore the application of computer vision and machine learning approaches to improve current methods used to investigate the perceptual mechanisms underlying face perception. For instance, one of these methods is to use eye-tracking devices. Such devices offer exceptionally rich information regarding direct attention. However, because the data provided by such devices is so rich, it is an effort in the scientific community to analyse it and recognise visual patterns to – possibly - investigate individual differences. And so, towards the end of this thesis, I develop new techniques for analysing eye-tracking data that will enable an improved understanding of the mechanisms of attention of human participants performing computer-based and 'in the wild' studies of face perception.

The following sections of the general introduction introduce the main background to my experimental work. First, I discuss how error-prone unfamiliar identity verification performance can be for humans and DCNNs before reviewing image manipulations that can further deteriorate performance. Second, I introduce potential routes for reducing such errors, focussing on the promise of 'fusion' approaches that combine independent decisions made by humans with high-face processing abilities and state-of-the-art DCNNs. These fusion effects rest on divergence in cognitive processing between humans and DCNNs, so it is important to understand individual differences in cognitive mechanisms behind face identity decisions. Third, I present evidence that the ocular strategy of humans sampling information from faces changes with their face-processing abilities, enabling a better understanding of the mechanisms behind individual differences and how these differ from DCNN processing. Finally, I will outline the experimental chapters and how these relate to the aims of this thesis.

Human performance in face identity processing tasks

Most people can effortlessly recognise and match the faces of familiar individuals. For example, when matching familiar face images by identity, people typically achieve ceiling-level accuracy even when images have poor lighting, different angles of view, poor image quality, and 'disguises' (i.e. hats, glasses, and beards)(Burton, Wilson, Cowan, & Bruce, 1999; Hancock, Bruce, & Burton, 2000). However, when performing these same tasks with unfamiliar faces, performance decreases significantly. Even in a simple matching task, where participants must decide if two face images presented side-by-side are of the same person or different people, the observed accuracy is around 80% (Burton et al., 2010).

Figure 1.1 shows a pairwise face-matching decision from a standardised test of unfamiliar face-matching ability created by Burton and colleagues (2010), the Glasgow Face Matching Task (GFMT). These images are taken using two different good-quality cameras in a controlled environment minutes apart in time. Therefore, in this task, the face images are aligned, with no discrepancies in lighting conditions, angles of view, quality or age-related appearance. And yet people are highly susceptible to errors on this task. Half of the participants get this particular pair of Figure 1.1 wrong, and the average person makes 20% of mistakes considering the entire test of 40 image trials. In more challenging tests, using not necessarily perfect images, the overall accuracy can drop up to 40% (e.g. Davis & Valentine, 2009; Henderson, Bruce, & Burton, 2001; Phillips, Yates, et al., 2018). Notably, this reduction in accuracy automatically questions the validity of using humans to perform such tasks in security settings where the task is to compare -for example- the identities of unfamiliar people captured by surveillance CCTV footage.

Figure 1.1. Example of a face-matching trial of the GFMT (Burton et al., 2010). Do these two images show the same person or different people? Around half of the people incorrectly respond that these images are of the same individual.

How face images affect unfamiliar face matching performance in applied settings

Face images vary in a range of factors. Some of these could be related to imaging conditions, for example, the quality of the image, illumination, angles of view, etc. Others are related to the face itself, for example, appearance changes due to aging, make-up, facial hair etc. Combined, these sources of variation can cause the task of unfamiliar face matching in realistic conditions to be far more complicated than the example shown in Figure 1.1.

Jenkins and colleagues (2011) demonstrated this by creating a task where participants sorted a set of 40 'ambient' images by identity. In this task, they gave participants 40 pictures containing realistic variations in illumination, angles of view, quality, appearance, etc. When the identities were of unfamiliar individuals, the average participant sorted the 40 images into seven identity piles, despite the correct answer being that there were just two identities equally distributed in the set. However, participants could easily find the two identities when the images were of familiar individuals. This result demonstrates that our ability to recognise unfamiliar individuals is susceptible to natural changes in appearance and external factors (i.e. image quality, angles of view, illumination, etc.)(Jenkins, White, Van Montfort, & Burton, 2011).

The quality of an image is another essential factor in provoking different outcomes regarding face identity decisions. This factor has applied importance in security and forensic settings where it is often necessary to compare the identity of a suspect depicted on a high-quality mugshot photograph against those taken by – for example – CCTV surveillance

systems. Even high-quality modern CCTV cameras are known to introduce distortions, artefacts, and unusual viewing angles due to being positioned far away (i.e. on rooftops and ceilings) from the subject of interest (see Seckiner et al., 2018 for a review). This disparity in the nature of mugshot versus CCTV photographs is an important issue because the distinct sources, combined with the unfamiliarity of the suspect, can further induce more errors, which could lead to the arrest of an innocent person.

Image quality has significantly impacted face-matching performance within lab-based performance testing. For example, in developing the Glasgow Face Matching Test 2 (GFMT2), White and colleagues (2022) included variations in head angle, expression and subject-to-camera distance. Notably, these sources of image variance were not present in the original GFMT. So the GFMT2 was designed to be more challenging than its predecessor and more representative of the difficulties encountered in applied face-matching tasks. When one of the face images in a pair was pictured at a distance, thereby reducing the image quality (see Figure 1.2), participants were 10% less accurate on average compared to when two images were of similar quality.

Figure 1.2. Example of a face-matching trial of the GFMT2 (White et al., 2022). Do these two images show the same person or different people? Despite the task being similar to the one shown in Figure 1.1, the discrepancies in image quality make the task more challenging, but also more representative of the difficulty of face-matching decisions in important applied settings.

In applied tasks, for example identifying culprits from CCTV, the task of unfamiliar face matching is highly error-prone. In studies using realistic CCTV quality images, and where there are changes in appearance of individuals between CCTV and mugshot images, errors can be from 25% to almost 50% (e.g. Davis & Valentine, 2009; Henderson, Bruce, & Burton, 2001). In a study by Davis and Valentine (2009), participants had to match the

identity of people present in a room against old video footage (i.e. 1-year-old) taken from a middle-range distance. This study shows that 25% of decisions were errors. Interestingly, adding disguises to the videos (i.e. hats, glasses, etc.) increased identification errors from 25% to 48%. As another example, in more straightforward tasks (i.e. a 1-to-1 matching task), Henderson and colleagues (2001) show that 55% of their participants wrongfully decided that the video footage of a robber did not match his good-quality photograph. Crucially, even though these values reveal some significance, it is important to notice that the entire forensic task in real-world situations requires a sequence of distinct actions that aid the investigation beyond matching one identity to another. To illustrate, we could consider details such as the location, timing, a person's walking style, individuals who witnessed the event, and other forensic methods to guide these choices. Still, because identity-matching is an important aspect of the overall task, it is critical to improve the mechanisms for decisions in this area.

Individual differences in identity processing

The previous section showed that variation in unfamiliar face-matching performance is affected by factors relating to the stimulus. But it is important to address that factors relating to the observer can also affect the outcome of the decision. As an example, researchers show that working under time pressure (Fysh & Bindemann, 2017), sleep restriction (Beattie, Walsh, McLaren, Biello, & White, 2016), and high anxiety levels (Attwood, Penton-Voak, Burton & Munafò, 2013) can also significantly affect the outcome of facial identification decisions.

But while these factors relate to transient variations in participants' mental states, viewers also vary in their intrinsic ability levels. It is now well known that face identity processing ability differs substantially in the population (see White & Burton, 2022 for a review). And while people generally show poor performance in unfamiliar face-processing tasks, there is also an extensive range of variation in this task inter and intra-individuals for face processing paradigms (see Bobak et al., 2023 for review). Interestingly, however, studies also show that this ability is consistent over time, showing high test re-test correlations (e.g. Sutherland et al., 2020; Germine et al., 2015, Balsdon et al., 2018; White et al., 2021) and generalisation across different types of face identity processing tasks (e.g.

McCaffery, Robertson, Young, & Burton, 2018). In addition, this ability appears to be hereditary (Wilmer et al., 2010; Zhu et al., 2010; Shakeshaft & Plomin, 2015). This ability spectrum ranges from those who show severe impairments in recognising or matching faces (e.g. Duchaine, Wendt, New, & Kulomäki, 2003) - sometimes not even recognising themselves in a mirror – to those with outstanding face identity processing performance in multiple standard deviations above the mean (e.g. see Dunn, Summersby, Towler, Davis, & White, 2020).

The literature distinguishes these individuals possessing different face-processing performances into three primary performance groups, from lower to extreme face-processing abilities: Prosopagnosics, Typical viewers (i.e. average population), and Super-recognisers. People at the lower end of the performance spectrum include individuals diagnosed with Prosopagnosia (Greek: *prosōpon,* face; *agnōsia*, ignorance). Prosopagnosia is a neuropsychological disorder that could be acquired by trauma (Acquired Prosopagnosia, see Damasio, Damasio, & Van Hoesen, 1982) or due to intrinsic hereditary and natural developmental factors (Developmental Prosopagnosia, see Behrmann & Avidan, 2005). Impaired face processing in prosopagnosia is not due to impairment in broader intellectual or low-level visual functions (Susilo & Duchaine, 2013).

In contrast, the other extreme side of the ability spectrum comprises individuals possessing outstanding performance in both unfamiliar recognition and matching tasks, called super-recognisers (Russell et al., 2009). Super-recognisers can outperform groups of typical viewers in face-processing tasks despite the cognitive and perceptual underpinnings that explain their superiority remaining unclear. Still, super-recognisers are said to be a solution to real-life applied settings involving processing facial identities. Being used by, for example, police forces in the UK (see Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016) or other countries[1].

Individual differences in face processing ability are robust to changes in task format. For instance, unfamiliar face-matching demands comparing the identity information between two different face images without requiring any memory component. On the other hand, face-memory tasks demand memorising identities to be posteriorly recognised. It is plausible to assume that these apparently different tasks rely on different mechanisms.

[1] https://www.superrecognisers.com/post/new-super-recogniser-contract-with-europe-s-largest-police-force-bundespolizei-signed-in-greenwich

However, studies report high correlations between face memory and matching abilities in the range of 0.5 to 0.7 (Verhallen et al., 2017; Balsdon, Summersby, Kemp, & White, 2018; McCaffery, Robertson, Young, & Burton, 2018), showing that there is a generalizable component of face identity processing tasks.

In addition to this 'convergent validity' of face identity processing tasks (see Wilmer et al. 2012), studies also show that face processing performance shows weak associations with other visual processing tasks (e.g. general object recognition). To illustrate, studies report that face-recognition ability, measured by the Cambridge Face Memory Test (CFMT)(Duchaine & Nakayama, 2006), shows weak correlations with unfamiliar abstract art recognition (r=0.26, Wilmer et al., 2010; r=0.26, Wilmer et al., 2012) and car recognition (r=0.29, Shakeshaft & Plomin, 2015; r=0.37, Dennett et al., 2012). In addition, other studies report that face-matching ability, measured by the GFMT, shows weak correlations with fingerprint matching (r=0.18), firearm matching (r=0.20), and artificial 'potato print' matching (r=0.41)(see Growns, Dunn, Mattijssen, Quigley-McBride, & Towler, 2022). While demonstrations of convergent validity show face identity processing ability is generalizable across task format, weak associations with object memory show 'divergent validity', in that face processing appears as its own, somewhat isolated, ability (see also Richler et al., 2019; and White & Burton, 2022 for a review).

Improving human performance

It is well established that standard participant cohorts of university students perform poorly on unfamiliar face-matching tasks. More concerningly, professional staff who perform unfamiliar face-matching in their daily work show comparably poor levels of accuracy (see White, Towler, & Kemp, 2021 for a review). It is, therefore, natural to ask whether anything can be done to improve the accuracy of face-matching decisions in applied settings, especially in high-stakes security and forensic tasks. One possible – and relatively straightforward - solution relies on the scientific study of individual differences in face identity processing ability, which has been reviewed in the previous section. That is, selecting people that score highly on face identity processing tests has been proposed as a possible solution to this problem (e.g. White, Kemp, Jenkins, Matheson, & Burton, 2014; Bobak, Dowsett, & Bate, 2016).

Due to their natural and stable face-processing ability, super-recognisers have increasingly received more interest in policing and national security surveillance tasks as a pragmatic solution to avoid errors related to identity recognition. As an example, the London Metropolitan Police have been selecting individuals based on their face-processing abilities to be part of their team (Davis, Forrest, Treml, & Jansari, 2018; Davis, Lander, Evans, & Jansari, 2016; Robertson, Noyes, Dowsett, Jenkins,& Burton, 2016), as well as other police forces outside the UK (e.g. German Federal Police[1])(see Ramon, Bobak, & White, 2019). However, despite the use of super-recognisers for such roles in security and forensic settings, studies supporting this solution primarily focus on correlations between performances on laboratory-based tests (e.g. Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Balsdon, Summersby, Kemp, & White, 2018). Whether these tests are able to predict accuracy in everyday police work is not tested (see Ramon, Bobak, & White, 2019).

An alternative solution is to apply training procedures to improve face-processing abilities. It is typical in professional organisations that require staff to make face identity decisions – for example, passport issuance officers (White et al., 2014) – to provide training. But is this training effective in improving unfamiliar face-matching accuracy? To answer this question, Towler and colleagues (2019) systematically investigated the content of eleven different professional training courses. They also tested the effectiveness of four of these courses by comparing face identification accuracy in large groups of participants before and after they had completed the training. They report that short courses, such as the ones frequently used by government agencies, do not improve facial identification accuracy. On the other hand, they provide evidence that a more prolonged course (i.e. a 3-day training course) significantly improved the facial identification performance of participants, but only on some of the tests. Thus, it is still unclear if training courses can directly improve facial recognition performance, and where it is effective it is likely to rely on the length of training processes (Towler et al., 2019).

Relatedly, recent studies have shown that forensic examiners (i.e. professionals who perform the investigation in identification procedures) outperform standard accuracy on unfamiliar face identification tasks (see White, Towler, & Kemp, 2021 for a review). Forensic examiners are trained personnel who aid in providing evidence in court proceedings (e.g. Jain, Klare, & Park, 2012; Dessimoz & Champod, 2008; Jain & Ross, 2015). Studies show that forensic examiners exceed the normative levels of face identification performance

compared to typical viewers or students (White, Phillips, Hahn, Hill, & O'Toole, 2015) and show comparable accuracy to super recognisers in unfamiliar face matching tasks (Phillips et al., 2018). One possible explanation for their high accuracy of unfamiliar face matching is due to the training they receive in close feature-based comparison of face images [2]. However, this argument conflicts with previous studies investigating the effectiveness of training procedures (e.g. Towler et al., 2019). Alternatively, examiners may possess inherently higher face identity processing ability, and they gravitate to the examiner roles through 'self-selection'. However, this is not consistent with apparent qualitative differences in the way that examiners perform the task, which dissociates them from typical viewers and super-recognisers (see White et al., 2015). It appears likely then that a combination of training, prolonged practice, mentorship and experience lead to superior abilities of examiners, but the underpinnings of such improvement remain unclear (see also Towler, Kemp, & White, 2021).

Technological alternatives

Another approach to overcoming humans' error-prone unfamiliar facial identification ability is to replace human decisions with automatic systems. Recent advances in technology, such as the multi-layer architecture of Deep Convolutional Neural Networks (DCNNs), mean that artificial systems can now achieve accurate face identification across a wide range of image variations (e.g. expression, illumination, angle of view, etc.). Recent DCNNs, when properly trained for identity recognition, can achieve comparable performance to the most accurate human participants: super-recognisers and forensic facial examiners (see Grother, Ngan & Hanaoka, 2019; Phillips et al., 2018). DCNNs are trained via backpropagation to associate face images with identity labels. The robustness of DCNNs producing accurate identity recognition even with diverse image variation is a notable property of their performance compared to previous generations of face recognition algorithms. This is most likely due to the massive databases of images captured in unconstrained environments that are used to train these neural networks (Phillips, 2017; O'Toole et al., 2018).

[2] Facial Identification Scientific Working Group. 2011. Guidelines and recommendations for facial comparison training to competency. See www.fiswg.org/document/.

A schematic illustration of the computational process involved in matching face images using a DCNN is shown in Figure 1.3. DCNNs represent the identity of a face image as a relatively compact numerical description – i.e. a feature vector – that can be used as a quick tool to compare the identity information of two previously unseen faces (Figure 1.3, top panel). Feature vectors are derived through a hierarchical process of progressive abstraction from the original pixel content of the image via convolution and pooling operations (Figure 1.3, top left). Weightings of neurons in the network are adjusted during training, where millions of images are fed to the network, which learns to associate these images with identity labels via backpropagation. Feature vectors of trained networks can then be used to verify the identity of unfamiliar faces by projecting the vectors of two images into the multidimensional 'face space' defined by the feature vectors and measuring the distance between them in the space (Figure 1.3, top right).

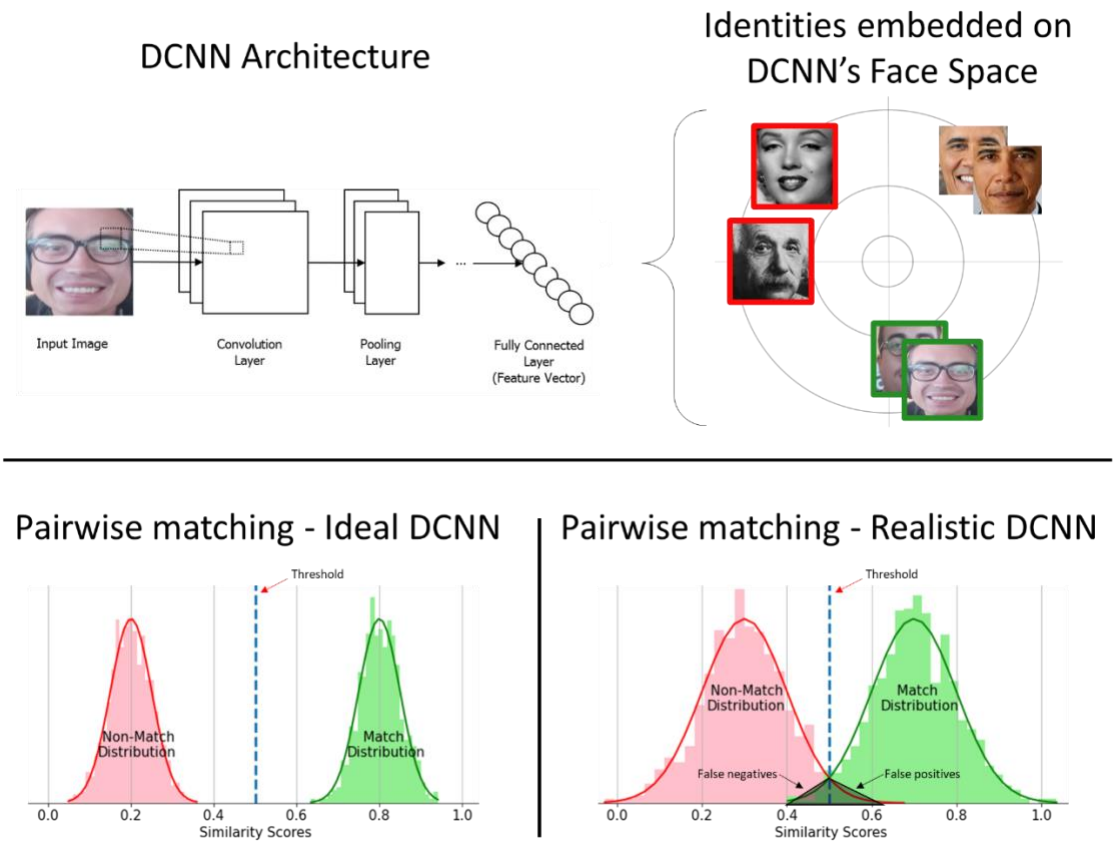


Figure 1.3. Schematic diagram showing how DCNNs process face identity. See text for details.

In an appropriately trained model, similar identities are clustered more closely together in face space, and different identities are physically further apart. At the bottom panel of Figure 1.3, we show the hypothetical distribution of Euclidean distances between

images of non-matching identities in red and matching identities in green. By applying a simple threshold between these distributions, the algorithm can decide if the images are of the same person or different people. However, despite DCNNs showing the potential to discriminate any two faces from matching or non-matching identities (see bottom left panel of Figure 1.3), in realistic scenarios, DCNNs produce errors observed as overlaps between such distributions, resulting in a range of similarity scores for which the DCNN can not make a definitive identity judgment (see bottom right panel of Figure 1.3). And so, while DCNNs perform very accurately – as well as the best performing human participants (Phillips et al. 2018) – they continue to make errors on challenging tasks.

There are other reasons why entrusting face identification decisions to DCNNs without human supervision can be problematic. For example, DCNNs are trained using face databases varying in demographic composition (e.g. in age, biological sex, ethnicity, etc.). Because DCNNs optimisation is essentially a 'statistical fit' to any given database, the relative accuracy of algorithms with particular demographics is sensitive to the demographic composition of the training databases (Vera-Rodriguez et al., 2019). This effect, notably, has caused biases in facial recognition DCNNs whereby they show lower accuracy for minorities compared to majorities, raising possible ethical issues (e.g. different races: Cavazos, Phillips, Castillo, O'Toole, 2020). Furthermore, it is important to address that the ethical dilemmas stemming from discriminatory patterns when AI is used within decision-making paradigms have ignited extensive discussions among the scientific community (see Birhane, 2021 for a review).

Another potential problem with DCNNs is their apparent sensitivity to variations in image quality since their training input is typically composed of high-quality images (Vogelsang et al., 2018). This problem has a significant impact because a contentious use-case of this technology is in police investigation, where images of poor quality (e.g. CCTV images) are commonplace. Early work showed DCNNs have lower object classification accuracy for lower-quality images (e.g. Dodge & Karam, 2016). It is perhaps not surprising then that using such technology for surveillance, without any level of human supervision, has caused the arrest of innocent people (e.g. see Williams, 2020, for a report where a man was wrongfully arrested purely because of algorithms' decisions).

Human-AI hybrid approaches

As outlined in the previous section, current facial recognition technology is not currently accurate enough to operate completely independently of human oversight in most applied settings. In fact, this problem is acknowledged in government policy across an extensive range of domains where AI is being used to make government decisions, as legislation in many countries requires human oversight of algorithmic decisions (e.g. Green, 2022). But again, as we have already seen, human performance on these tasks is also error-prone, leading to a circular problem.

Because AI cannot operate completely independently for identity verification in security proceedings, it is necessary to have human reviewers. However, studies have shown that human review is – itself - error-prone, with both standard participant groups and people who review FR technology in their daily work showing 50% error rates (White, Dunn, Schmid, & Kemp, 2015). One solution to this problem is to utilise people with higher accuracy than standard participant groups in this role, such as super-recogniser and forensic facial examiners. Moreover, a crucial requirement arises for a complete absence of decision sharing between humans and DCNNs. This is due to the observed inclination of humans to disregard their decisions and completely bypass the algorithm's decisions when provided access to its similarity scores - or binary judgments (e.g. Fysh & Bindemann, 2018; Howard, Rabbitt, & Sirotin, 2020; Carragher & Hancock, 2023).

Recent research points to an alternative way that human and DCNN decisions can be combined to improve accuracy in applied settings. Studies have investigated the benefit of 'fusing' independent judgments made by humans and DCNNs. Phillips and colleagues (2018) illustrated that making a pairwise combination between independent judgements made by human forensic facial examination experts and state-of-the-art algorithms (i.e. DCNNs) improved the quality of face identification decisions to ceiling levels. Importantly, this accuracy was higher than either humans or algorithms could achieve alone or in human-human or DCNN-DCNN pairs (see also Knoche and Rigoll, 2023).

Prior work has shown that aggregating (i.e. averaging) decisions made by groups of individual human participants can significantly improve the accuracy of decisions compared to when using individual participants alone (White et al., 2013; Jeckeln et al., 2018). This effect leading to improved performance by aggregating responses is commonly known as

the 'Wisdom of Crowd'. But the additional boost of teaming DCNNs and humans together in these joint decisions is potentially even more interesting. One proposal is that this result is due to the increased diversity in the cognitive strategies employed by humans and DCNNs compared to two human participants (Towler et al., under review; Hong & Page, 2004). And so assuming that DCNNs and human high-performers use different processes to reach face identity decisions, the wisdom of the crowd effect caused by combining humans and DCNNs could lead to even more accurate facial identification decisions (see Kittler, Hatef, Duin & Matas, 1998; O'Toole, Abdi, Jiang, & Phillips, 2007; White, Burton, Kemp & Jenkins, 2013; Hu et al., 2017; Jeckeln et al., 2018).

Analysing the wisdom of crowd effects produced by combining humans and DCNNs can help understand the similarity/differences between underlying cognitive processes in humans and algorithms (O'Toole, Abdi, Jiang, & Phillips, 2007). In turn, a better understanding of how their processing diverges can identify opportunities to improve such human-AI hybrid systems. In addition, because we know there are significant differences in how individual humans process faces – and the accuracy they attain – there is a considerable knowledge gap in understanding how best to merge individual humans with DCNNs' decisions. Furthermore, there are still gaps in the literature regarding how to establish a proper 'cognitive' measurement between humans and DCNNs and how they process faces (but see Hill, Roodenrys, & Clifford, 2019). As I will describe in the next section, this research agenda can also benefit theoretical understanding.

Using DCNNs as a model of human processing: comparing man vs machine

There are commonalities and differences between the cognitive processes that DCNNs and humans engage in when processing faces—showing that aggregating DCNNs and human inputs illustrate differences in their cognitive processes. But at this point, there is only an emerging theoretical understanding of these differences and how they affect the performance of human-AI systems.

One way to compare face processing in humans and DCNNs is by examining how changes in input affect their relative performances. For example, a standard result in the study of face identity processing in humans is that people are better at recognizing faces from their own ethnic group. This is known as the other-race effect (see Meissner &

Brigham, 2001), and pre-DCNN face recognition algorithms were also shown to display this effect (Furl, Phillips, & O'Toole, 2002). In the study by Furl and colleagues (2002), face recognition algorithms developed in Western countries were more accurate at identifying Western faces relative to East Asian faces and the opposite pattern was found for algorithms developed in East Asia. Similar demographic biases are known to exist in DCNNs (Grother, Ngan, & Hanaoka, 2019; Cavazos, Phillips, Castillo, & O'Toole, 2020), possibly being caused by 'unbalanced' training databases (i.e. not containing compared number of images per race, gender, etc.) (Vera-Rodriguez at al., 2019).

Another standard effect in psychology that has recently been replicated in DCNNs is the 'caricature effect', where caricature images are recognised faster than standard images (see Rhodes, Brennan, & Carey, 1987). An interesting study by Hill and colleagues (2018) shows that DCNNs accuracy increased when matching caricatures compared to standard face images. This improvement might be because DCNNs represent the identity information of caricatures as more distinct than others, reducing the probability of confusion with another identity (Hill et al., 2018). Critically, these pieces of evidence point to the possibility that the 'face-space' encoded in DCNNs is similar to the latent face space used in human judgments (see Valentine, 1991).

Another approach for understanding similarities between humans and DCNNs' face-space is directly correlating their ratings of similarity between two face images. Towler and colleagues (under review) show item-level correlations between face similarity ratings by super-recognisers (SR), forensic examiners (FE) and 7 DCNNs performing a face identity matching task. Average item-level correlations of similarity ratings were relatively medium for the two human expert groups (matching face image pairs $\rho = 0.25$; non-matching pairs $\rho = 0.37$). The seven DCNNs, on the other hand, showed high agreement in similarity ratings ($\rho = 0.60$). Most importantly, DCNNs and humans showed *disagreement (i.e. negative correlation)* on ranking the similarities between face pairs when faces were of different people (matching face image pairs $\rho = 0.27$; non-matching pairs $\rho = -0.19$). This shows a potential divergence in how humans and algorithms assess similarities between different faces. In addition, it may suggest that the face-space of humans and DCNNs are structured differently.

This result contrasts with other recent studies showing stronger correlations between DCNN and human face similarity ratings. In one study by Grossman and colleagues

(2019). Their study correlated the similarities between human intracranial EEG signals and multiple layers of a DCNN (VGG-16: Simonyan and Zisserman, 2014) while processing faces. They show that the geometrical distances measured by iEEG signals strongly correlated (e.g. $\rho > 0.6$) with the geometrical distances between activations in some of the middle layers of the DCNN when observing two different images (Grossman et al., 2019). This result shows remarkable similarities between a possible face-space configuration of humans and DCNNs, providing evidence that they might structure their facial identity representation similarly. Other studies using human similarity ratings have found similar results (Jozwik et al., 2022). The extent of divergence between human and DCNN face processing, therefore, remains unresolved.

In this thesis, I will investigate the effect of decrements in image quality on humans and DCNNs, to ascertain whether this degrades human and algorithm performance similarly. As discussed, separate studies have shown humans and algorithms suffer from decrements when image quality is reduced, but we do not know of any study that directly compares this effect. This comparison will provide practical guidance on the use of DCNNs and humans in tasks where image quality is poor, and also provide a converging line of evidence to inform whether DCNNs and humans use similar face identity processing.

Using Eye-tracking to understand processing differences underlying face identity processing in humans

While one focus of this thesis is on understanding the processing differences between DCNNs and humans, another aim is to improve understanding of the processing differences between individual humans that give rise to large individual differences in behaviour. An essential focus of prior work on this topic has been to understand how face perception relies on the 'holistic' processing of faces as unitary gestalts rather than feature-by-feature part-based analyses (Farah, Wilson, Drain & Tanaka, 1998; Richler & Gauthier, 2014). The gestalt view proposes that the mechanism to process faces somehow compresses all facial features and their information in a single 'holistic' variable to represent someone's identity.

Initial work aiming to understand the underlying processing differences that give rise to individual differences in ability has focused on the dichotomy between holistic and 'part-

based' processing. The literature suggests three main image manipulations to illustrate that faces are perceived holistically: The Composite Face Effect (Young, Hellawell, & Hay, 1987); The Face Inversion Effect (Yin, 1969); and the Part-Whole effect (Tanaka & Farah, 1993). To create composite faces, researchers mix the top half of a face with the bottom half of a different face, creating a 'new' identity based on two distinct individuals. When these face halves are aligned, the manipulation disrupts participants' ability to recognise the identity of the face halves. When the halves are not aligned, participants are able to recognise the source identities (see Young, Hellawell, & Hay, 1987). The disruption in identity recognition caused by aligned face halves shows that faces are perceived holistically and not as a sum of decomposable facial features.

The Face Inversion effect refers to the greater impairments in identity recognition when faces are turned upside down compared to traditional when other classes of objects are turned upside down (Yin, 1969). This is thought to result from inversion forcing a more 'part-based' approach due to disruption in holistic processing (Carey & Diamond, 1977). Lastly, the part-whole effect (Tanaka & Farah, 1993) shows improved memory for single face features (e.g. the eyes) when presented in the context of a full face compared to when presented alone (Tanaka & Farah, 1993). This reduction in performance is also argued to illustrate that facial features are represented in the context of the whole face rather than as isolated 'part-based' facial features.

However, despite these findings suggesting that a 'holistic' representation enables identity processing, studies found that the holistic processing measures themselves do not predict individual differences in identity processing ability. For example, Rezlescu and colleagues (2017) measured the associations between face recognition ability and the aforementioned holistic processing measures. Their work shows that only the inversion effect could predict face processing ability. In addition, they report that the measures used for holistic processing do not correlate with themselves. This result is interesting because it shows that, despite measuring what seems to be a common holistic processing mechanism, the measures may, in fact, be measuring different processes (see also Sunday, Richler, & Gauthier, 2017).

Another, arguably more direct method for measuring the qualitative aspects underlying face processing ability is to measure individual observers' eye movements when they perform face identity processing tasks. Some studies using eye-tracking devices found

that the central area of the face (i.e. the nose) is the more functional area to fixate for identity recognition (e.g. Hsiao & Cottrell, 2008). This appears to show that humans can focus on a single position and still extract enough identity information for accurate recognition. The argument is that focusing on the face centre would allow them to extract all facial features by using the surrounding areas of their fixation, corroborating the holistic approach for face processing. However, in another study by Henderson and colleagues (2005), researchers compared the accuracy of participants recognising faces under two conditions, where they could freely investigate facial features or keep their fixation steady. Their result shows that recognition accuracy was 28% higher for the condition where participants freely moved their eyes. And so, while holistic processing may be important in face identification, it is clear that visual exploration of facial features is also important for recognition.

Human eye movement patterns when processing faces varies depending on an individual's face-processing ability. Recent studies have found that super-recognisers tend to significantly sample more information from the central area of the face (Bobak et al., 2017; Bennetts, Mole & Bate, 2017) compared to typical viewers when observing scenes. This result corroborates the idea that how humans sample information from faces might reflect traces of individual differences in face-processing ability. That is, the results from Bobak and colleagues (2017) illustrate that super-recognisers might show improved holistic processing compared to typical viewers due to their higher sampling from the face centre. However, this result contrasts with other evidence showing that super-recognisers require less visual information to keep accurate than typical viewers (Royer, Blais, Gosselin, Duncan & Fiset, 2015). Using the Bubbles technique (Gosselin & Schyns, 2001), researchers can control the amount of facial information participants can observe in a given image through transparent 'bubbles' of various aperture sizes. When processing faces through such a technique, Royer and colleagues (2015) provide evidence that super recognisers require less information (i.e. fewer bubbles) compared to typical viewers to keep accurate recognition (Royer, Blais, Gosselin, Duncan & Fiset, 2015). And so, studies remain a somewhat mixed picture of the fundamental differences in information sampling that underpin differences in face identification ability.

In this thesis, we propose investigating the differences in eye-movement strategies leading to superior performance in face recognition and matching tasks. In addition, such

understanding of information used for face processing could be used to -for example- improve face recognition technology or even training procedures to improve the quality of decisions in security settings.

Lab-based eye tracking research and ecological validity reflecting human face processing

Another aim of this thesis is to broaden the study of human eye-tracking studies by incorporating recent advances in computer vision with wearable eye-trackers. Decades of research have studied socially-directed attention by analysing people's eye movements as they view images of faces – or people - presented on computer screens (e.g. the classic study of Yarbus (1967)). However, one big issue related to research on this topic is that photographs of social scenes - inevitably - do not represent the multidimensional and dynamic reality of our social experiences. Some studies found different fixation patterns when participants engaged in real face-to-face interactions compared to a similar task but in screen-based stimuli (e.g. Nasiopoulos, Risko & Kingstone, 2015). Thus, this difference in ocular patterns provoked by different social contexts indicates that computerised laboratory tasks might be inadequate to capture real-world social attention (see also Kingstone, 2009).

Taking social attention research out of the lab may also benefit understanding individual differences in face processing. Recent lab-based eye-tracking studies have shown large individual differences in how people attend to social scenes shown on screens. These studies also point to a genetic basis underlying people's social attention (Constantino et al., 2017; Kennedy et al., 2017). Other studies have found significant and stable individual differences in people's face-processing abilities (White & Burton, 2022), and -as previously discussed- such results are associated with different patterns of eye movements to faces and people in lab-based tasks (Bobak et al., 2017). Whether these patterns hold up in natural settings – where 'social stimuli' are real people – remains unclear.

There is existing literature investigating social attention 'in the wild' while participants navigate real-world ambients wearing an eye-tracking device (e.g. Foulsham, 2020; De Lillo et al., 2021). These devices enable researchers to study social attention to faces and person perception *in situ*. However, this requires experimenters to manually manipulate long video recordings and code what is being fixated for every video frame. As a result, even coding simple aspects of gaze fixations - such as counting person fixations vs

non-person fixations during any task - would be highly time-consuming (see Hessels et al., 2020). Thus, examining social attention in naturalistic environments using more extensive samples of participants is impractical at the resolution afforded by such devices. Therefore, in the final chapter of this thesis, we aim to develop an automated solution to investigate social attention 'in the wild' using wearable eye-tracking devices using automatic person and face detection artificial intelligence algorithms (OpenPose: Cao et al., 2019).

Thesis aims

Following this introduction, Chapters 2 and 3 will investigate the strengths and weaknesses of humans and artificial intelligence in applied settings in making decisions regarding identities. Chapters 4 and 5 then use artificial intelligence as tools to explore the underpinnings behind individual differences in face processing ability. Chapter 6 provides a general discussion of the findings. Importantly, all analysis code and data supporting our findings and conclusions are available from the authors.

The aim of Chapter 2 is to provide insights into differences between humans and state-of-the-art facial recognition algorithms (i.e. DCNNs) processing images of unfamiliar individuals. In real-world tasks, the quality of images is uncontrolled, and in many important forensic and security uses of facial recognition technology, images are poor quality - for example - CCTV images. For this reason, understanding how humans and DCNNs face identity varied as a function of image quality is of practical importance. But while studies have looked at the impact of image quality on humans and DCNNs separately, their relative performance has not been directly compared. In Chapter 2, we conducted four studies comparing humans and nine different algorithms matching identities while manipulating image quality. This improves understanding of the strengths and weaknesses of humans and DCNNs in applied settings and points to cognitive processing differences between them.

The aim of Chapter 3 is to improve the understanding of how independent face-matching decisions made by humans and DCNNs can be optimally combined in hybrid human-AI facial recognition systems. The promise of using human-AI hybrid systems is shown by recent papers reporting response fusion approaches that improve conjoint human-AI face identity decisions (e.g. Phillips et al., 2018; Towler et al., under review). However, these studies only report findings using high-performing human participants

(super-recognisers and forensic facial examiners). And so, it is unclear whether combining DCNN and human decisions would operate as effectively when human participants are recruited from across the ability spectrum - as would be the case in most applied settings. Thus, in Chapter 3, we conduct computational studies to investigate how the relative accuracies of DCNNs and individual human participants affect the benefits of response fusion approaches. This provides a practical guide on how fusing human and DCNN decisions could operate optimally with humans and DCNNs of varying face identity processing abilities.

Chapter 4 uses eye-tracking technology to understand better the information sampling strategies used by typical viewers and 'super-recognisers' when matching and recognising faces. Prior work has shown that super-recognisers use different facial information to code identity information depicted in face images compared to typical viewers (Bobak et al., 2017; Bennetts, Mole & Bate, 2017; Royer et al., 2015). Here I use new analysis and methodological approaches founded on my background in computer science to understand the information sampled by high-performing participants and its computational value for face identification.

In Chapter 5, I again aim to make methodological advances in face perception research by incorporating state-of-the-art computer vision approaches. By combining face and body detection algorithms with wearable eye-tracking technology, I introduce a new automated method for measuring attention to faces and bodies when they navigate natural settings. This provides a new data source to compare with eye-tracking studies conducted while participants view faces on computer screens in lab-based tasks. Prior work has shown different patterns of attention to faces when comparing screen-based versus when viewing real faces with wearable eye-tracking technology (e.g. Nasiopoulos, Risko & Kingstone, 2015; Kingstone, 2009), but these studies have been confined to relatively constrained experimental scenarios. In Chapter 5, we measure attention to people and faces as participants walk on the university campus and engage in social interactions, using innovative approaches to process eye-tracking data incorporating artificial intelligence and analytic geometry.

Chapter 2 - Comparing human and machine performance when matching images of different quality

Introduction

Humans recognise and match images of familiar people with very high levels of accuracy. Interestingly, when faces are unfamiliar, overall performance decreases drastically (e.g. Burton et al., 2010). Even in a straightforward perceptual face-matching task - where face pairs are presented side-by-side to viewers who are required to make a binary match/non-match response - some pairs are misclassified by more than 50% of participants. For example, see the image pair shown in Figure 2.1.

Figure 2.1. Example of face matching pair used in the Glasgow Face Matching Task (GFMT) (Burton et al., 2010). The reader should judge if the face pair is of the same person or different people. 50% of participants incorrectly respond that the images are of two different people.

Average human accuracy on standardised unfamiliar face matching tests is around 80% (e.g. Bindemann, Avetisyan, & Blackwell, 2010; Burton et al., 2010; Bindemann, Avetisyan, & Rakow, 2012; Burton, White, & McNeill, 2010; Megreya, Bindemann, & Havard, 2011; Megreya, White, & Burton, 2011). Accuracy decreases further when matching lower-quality images (Collishaw & Hole, 2000; Goffaux & Rossion, 2006; Cheung, Richler, Palmeri & Gauthier, 2008; White et al., 2021). For example, recent work by White and colleagues (2021) measured how human accuracy decreased when processing a face-matching task containing images taken at a distance on a video camera compared to a high-quality digital SLR camera at a close distance. Average accuracy with low-quality images was roughly 10% lower compared to when both were high-quality faces.

This reduced accuracy for poor-quality images has important practical implications. CCTV system images record low-quality images characterised by low resolution, artefacts, unusual angles of view, poor lighting, and distortions (see Seckiner et al., 2018 for a review). As a result, the legal system often has to deal with situations where the evidence includes a facial image of poorer quality. Some studies have assessed the accuracy of participants recognising -and matching- unfamiliar identities depicted as CCTV images and found very poor matching accuracy. For example, Burton and colleagues (1999) assessed the accuracy of a group of experienced police officers recognising unfamiliar identities from CCTV videos (i.e. poor-quality videos) in high-quality images. They found very poor accuracy in both students and experienced police officers when participants were unfamiliar with faces.

Later, Davis and Valentine (2009) investigated the effect of matching unfamiliar faces in CCTV videos against a 'live' lineup of real individuals. In one experiment, participants had to match the identity of people present in a room against old video footage (i.e. 1-year-old) taken from a middle-range distance. This study shows that 25% of decisions were errors. Interestingly, adding disguises to the videos (i.e. hats, glasses, etc.) increased identification errors from 25% to 48%. In more straightforward tasks showing pairs of images simultaneously on a computer screen (i.e. a 1-to-1 matching task), Henderson and colleagues (2001) show that 55% of their participants wrongfully decided that the CCTV video footage of a robber did not match his good-quality photograph.

It is possible that accuracy would be increased when CCTV videos are viewed rather than static video frames. However, the dynamic information contained in the video does not appear to impart any advantage compared to static stimuli (Liu, Seetzen, Burton, &

Chaudhuri, 2003). Moreover, studies suggest that even years of experience performing facial image comparison from CCTV footage does not protect from matching errors to low-quality images (Lee et al., 2009; Norrell et al., 2015). However, a limitation of all of the studies previously mentioned is that they defined low image quality in qualitative rather than quantitative terms by manipulating quality. For example, Norrell and colleagues (2015) investigated the effects of Facial identity comparison between a high-quality 'reference' image against pictures taken from a CCTV camera using three different zooming settings – which they report as Quality 1, 2, and 3 - without any specification of intended levels of degradation of the image but its size in pixels. Therefore, these studies do not directly quantify the levels of image degradation and measure the effect on face-matching accuracy.

Apart from the practical implications, understanding the impact of image quality on face matching is of theoretical interest because it reveals the relation between accuracy and spatial frequency content of images (Costen, Parker & Craw,1994; Costen, Parker & Craw, 1996; Bindemann et al., 2013). Previous studies have measured the degree to which particular bands of spatial frequencies carry identity information. Importantly, to allow easy comparison between studies, they typically report these spatial frequencies in terms of cycles per face (cycles/face). Studies using low pass filters - such as Fast Fourier Transformations (FFT) and Gaussian filters – found that the face processing ability of humans is critically disrupted when presenting face images containing only information below eight cycles/face. There is no disruption when tuning the same filter to show only information below 16 cycles/face, indicating that the usable information for face recognition is probably within the range of 8-16 cycles/face (Costen, Parker & Craw, 1994; Costen, Parker & Craw, 1996). Subsequent work has narrowed this range to 8-13 cycles/face (Näsänen (1999); see Jeantet, Caharel, Schwan, Lighezzolo-Alnot, & Laprevote, 2018 for a review).

The evidence of reduced accuracy in processing lower-quality face images (i.e. blurred images) suggests that encoding identity information can involve information encoded in the higher spatial frequencies, which are absent in these images. However, as people remain able to identify faces in images even with substantial blurring levels (i.e. more than eight cycles/face; Costen, Parker & Craw, 1994; Costen, Parker & Craw, 1996; Näsänen, 1999). This appears to suggest that identity information within faces is carried

both in the spatial relation between facial features (i.e. the face configuration; see Collishaw & Hole, 2000) and in feature details carried by higher spatial frequencies.

A large body of evidence points to two 'channels' of information that can be employed in face processing – configural and featural information (see Collishaw & Hole, 2000; Bartlett, Searcy, & Abdi, 2003). For example, studies show evidence for the importance of configural information includes that face recognition is impaired by: changing the distance between features (e.g. Sergent, 1984); mixing face halves of different people to create the illusion of a new 'composite' face (e.g. Young, Hellawell, & Hay, 1987); and blurring faces, to exclude fine-grained information using low-pass filters or pixelated versions of images (e.g. Bachman, 1991; Costen, Parker, & Craw, 1994; Costen, Parker, & Craw, 1996). However, on the other hand, studies that present scrambled faces that disrupt the spatial relationship between them show above-chance performance, suggesting some contribution of feature detail (e.g. Tanaka & Farah, 1993; Bruyer & Coget, 1987) corroborating evidence that fine-grained featural information also carries identity information (see Goffaux, Hault, Michel, Vuong, & Rossion, 2005).

There is evidence that the relative contributions of configural and featural 'channels' of information vary as a function of the familiarity of faces, with a greater reliance on configural information for familiar face identity processing. For example, Burton and colleagues (1999) show that even when viewing familiar people in substantially blurred CCTV images, participants remained able to recognize them. In addition, Lobmaier and Mast (2007) show that participants could match sequentially presented images of unfamiliar faces by identity better when the features were scrambled compared to when they were blurred. They found the opposite pattern for familiar faces, suggesting that blurring images is especially detrimental to unfamiliar face matching, but less so when faces are familiar.

Artificial Intelligence is now a feasible alternative to human processing in applied settings, and so the impact of blurring on algorithm performance is also important to quantify. Inspired by the human brain, the multi-layer architecture of modern 'Deep Convolutional Neural Network' algorithms (DCNNs) allows them to distinguish thousands of different identities, representing a multi-dimensional "face space" encoded in the top layers of the network (O'Toole et al., 2018). After sufficient training, these top layers are powerful enough to capture identity information across a vast range of image variations (e.g. expression, illumination, pose, etc.) in a relatively compact numerical representation

(Hancock et al., 1996; O'Toole et al., 1999; Hong et al.,2016; Parde et al., 2017) that can determine – with high accuracy - whether or not two images of previously unencountered faces are of the same individual (see Figure 2.2). Some of the latest versions of these algorithms can achieve higher performance than the general human population when matching unfamiliar faces (see Grother, Ngan & Hanaoka, 2019 for a review; but see Phillips et al., 2018 and Blauch, Behrmann, & Plaut, 2020 for human-machine comparison).

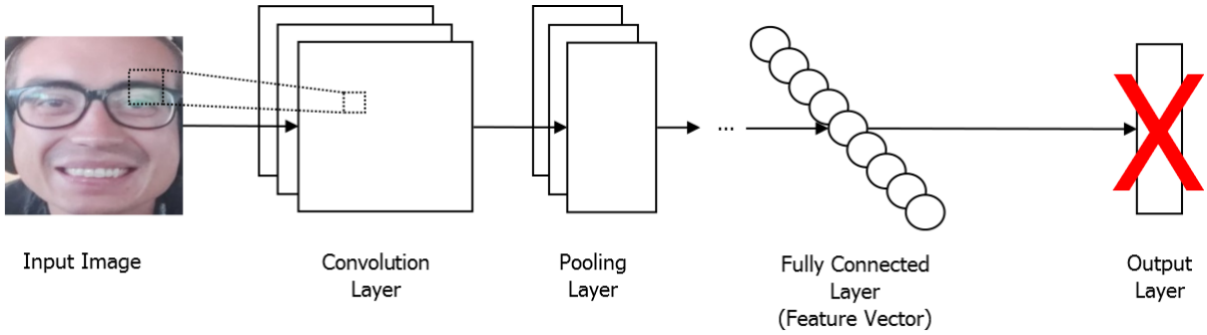


Figure 2.2. Schematic diagram showing the basic architecture of a Deep Convolutional Neural Network (DCNN). Input images pass through convolutional, pooling, and fully connected layers to activate a layer of identity labels in the output layer. During learning, associations are formed between representations in earlier layers and output layers, but these output layers are disconnected from the network after training is complete. The fully connected layer is then used to compare the identity of any two input images of faces that are 'unfamiliar' to the network. The fully connected layer is, therefore, a numerical representation of a face (feature vector) that can be used to determine if two images are of the same person or different people.

As with humans, one limitation of DCNNs is that they suffer considerably when presented with poor-quality images (e.g. see Dodge & Karam, 2016). For instance, Grother and colleagues (2019) reported that false-negative rates for modern vendor algorithms could be from 2 to 11 times higher when input images are from webcams or are unconstrained ("in the wild") images compared to high-quality "mugshot" frontal images. Some studies have employed visualisation techniques (e.g. t-SNE (Maaten & Hinton, 2008)) to understand how DCNNs aggregate identities. This approach suggests that poor-quality images are not represented in the same regions of face space as high-quality images of the same identity (O'Toole et al., 2018), thus, suggesting that the reduction in image quality also resulted in the loss of identity information in DCNNs. However, as these studies did not specifically manipulate image resolution in an unfamiliar face-matching task, the relationship between image resolution and face identification accuracy in DCNNs is not clear.

It is also unclear how the magnitude of image resolution-based declines in accuracy compares to that seen in humans. For example, is the performance advantage typically seen for DCNNS compared to human observers when tested with high-quality images maintained using reduced image quality? Some recent work suggests that algorithms may be more susceptible to degraded image quality because their training database comprises only relatively high-quality images (Vogelsang et al., 2018; Jang & Tong, 2021). This training is in contrast to humans, who initially learn to recognise the faces of their family members in infancy at a time when their visual input is blurred due to a lack of visual acuity (see Dobson, Teller, & Belgum, 1978; Rennels & Davis, 2008). Interestingly, the authors of these papers speculate that some differences between humans and DCNNs could account for this early learning stage with poorer quality inputs seen in humans. Interestingly, despite this speculation regarding the impact of image quality on recognition in humans and DCNNs, there do not appear to be any published studies that directly compare DCNNs and Humans processing blurred images.

This chapter investigates how the accuracy of face-matching decisions by humans and DCNNs is affected by changes in image quality. In applied face-matching tasks, for example, when using face images as forensic evidence, image quality can vary greatly from one face image to the next. And as aforementioned, both humans and DCNNs suffer considerably when presented with poor-quality images (e.g. see Dodge & Karam, 2016; White et al., 2021), but it is still unclear if they are affected to a similar degree. This question is of both theoretical and practical significance. Differences in the effects of image degradation on human and DCNN performance can point to differences in underlying processing. Such investigation could contribute to an emerging body of research that uses DCNNs as models of face processing in humans (e.g. Grossman et al., 2019) and also points to differences in the usability of humans and DCNNs providing identity verification for forensic analysis.

Experiment 1

Experiment 1 was designed to understand how varying levels of image quality affect face-matching accuracy in humans and DCNNs. Both humans and DCNNs matched face pairs that varied in image quality. We also manipulated whether both images in a pair were the same or different quality. This second condition has practical significance, for example,

simulating a comparison of a low-quality CCTV image with a high-quality reference image of a suspect. In addition, it is important to address that all studies in this thesis received ethics approval by the Human Research Ethics Committee at the University of New South Wales. All participants provided written informed consent prior to data collection.

Method

Participants

A total of 513 UNSW undergraduate students participated in the online study in exchange for course credits. Participants completed a face-matching task using one of two image filters to manipulate image quality: Fast Fourier Transformation (FFT) and Gaussian Blur (see Stimuli section for details). For the FFT versions, a total of 162 university students (41 male, 121 female; Age M= 19.43, SD= 2.3) participated in the Same-Resolution condition, while 168 (62 male, 106 female; Age M= 19.50, SD= 3.42) participated in the Different-Resolution condition. For the Gaussian blur versions, 100 university students participated in the Same-Resolution condition (37 male, 63 female; Age M= 19.16, SD = 2.11), and another group of 100 participated in the Different-Resolution condition (33 male, 66 female, 1 preferred not to answer; Age M= 19.48, SD= 3.52). We used large participant samples because each participant completed a subset of the experiment trials, which included six different levels of image quality for each of the conditions.

Deep Convolutional Neural Networks (DCNNs)

We used nine different DCNNs in this study. We collected those DCNNs from 5 different architectures trained on various datasets and implemented them using Keras (Chollet et al., 2015) or Pytorch (Paszke et al., 2019) in Python. See Table 1.1. All models were official models collected from the system developers (e.g. GitHub of the research in question). To ensure the replicability of our study, we used models from official sources so that other researchers could download the same DCNN parameters. DCNNs varied in architecture, training datasets, and deep learning libraries, thereby enabling the result of the study to be generalized across different DCNN systems. In addition, by using representative training sets, we aimed to provide a more realistic assessment of their capabilities in processing

facial data and making similarity judgments. We used an array of diverse DCNNs to draw a parallel to the experiences of the N individuals we tested, which originated from a variety of ethnical and geographical backgrounds. Notably, these varied experiences contribute to distinct perspectives among our human participants and DCNNs. This combination of multiple distinct viewpoints accentuates the depth and intricacy that form the foundation of our experimental approach.

DCNN	Architecture	Dataset	Python Library
1	ResNet50 (He, Zhang, Ren & Sun,2016)	VggFace2 (Cao et al., 2018)	Keras
2	ResNet34 (He, Zhang, Ren & Sun, 2016) (https://github.com/ageitgey/face_recognition)	VggFace (Simonyan and Zisserman, 2014); Face Scrub dataset (Ng & Winkler, 2014) and images from the internet (King, 2009)	Keras
3	ResNet50	VggFace2	Pytorch
4	ResNet50	MS-Celeb-1M dataset (Guo et al., 2016) fine-tuned on VggFace2	Pytorch
5	Se-ResNet50 (Hu, Shen, & Sun, 2018)	VggFace2	Pytorch
6	Se-ResNet50	MS-Celeb-1M dataset fine-tuned on VggFace2	Pytorch
7	VGG16 (Simonyan and Zisserman, 2014)	VggFace	Pytorch
8	FaceNet (Schroff, Kalenichenko, & Philbin, 2015)	VggFace2	Pytorch
9	Facenet	CASIA-WebFace (Yi, Lei, Liao & Li, 2014)	Pytorch

Table 1.1 Description of architectures, datasets, and Python libraries used in this study. We also enumerated DCNNs from 1 to 9.

Stimuli

We created the face-matching tasks in this study using the Glasgow Unfamiliar Face Database (GUFD) (Burton et al.,2010). The GUFD contains images of 303 identities photographed from various angles using two high-quality cameras (C1 and C2). We selected two frontal images of each identity from the database – one captured using C1 and one from C2. We excluded identities that did not have suitable frontal images from both C1 and C2. This selection resulted in a set of 602 frontal face images of 301 identities.

We used a DCNN to select 25 of the most challenging non-matching identity pairs. We achieved this by determining the 25 non-matching identity pairings rated as most similar by DCNN1 (see Table 1.1) amongst all identities in the database. Also, we employed DCNN1 to select the 25 most difficult match pairs from the remaining identities not included in the non-match pairings. Before matching, we used a Multi-task Cascaded Convolutional Neural Network (MTCNN) (Zhang, Zhang, Li, & Qiao, 2016) to detect, extract, and align the face across images. For five identities, the MTCNN failed to detect a face in one of the images (C1 or C2), so we excluded these five identities from the database, resulting in an image set of 296 identities. Next, DCNN1 received resized face images (224x224 pixels) as input, producing 602 feature vectors. We used the Euclidean distance between these feature vectors to infer the similarities between each face and all other identities. We summarised this processing framework in Figure 2.3.

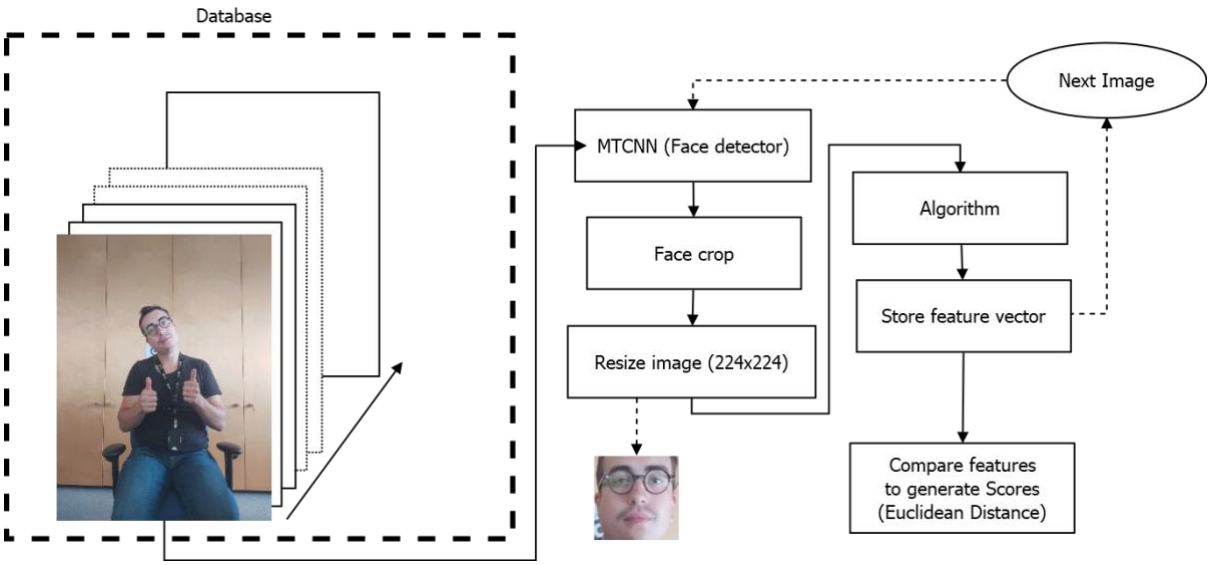


Figure 2.3. Description of the framework used to calculate similarities between faces in the dataset. We pre-processed faces using MTCNN (Zhang, Zhang, Li, & Qiao, 2016), then aligned and cropped

the images, resulting in square 224x224 pixels "face-only" images. We then used this image as the input for DCNNs to compare the Euclidean distance of its feature vector with other images in the dataset. Image taken for illustration purposes.

We created stimuli conditions using two distinct low-pass filters. We used both Fast Fourier Transformation (FFT) and Gaussian blurring filters to produce two different low-quality versions of each image. Because FFT filtering produces known visible image distortions (known as the "ringing-effect", see Gibbs, 1898, and an example of the ringing effect in the middle left image of Figure 2.4), we also employed a Gaussian filter version of the same stimuli, as Gaussian filtering does not produce distortions in the image (see bottom left of Figure 2.4). We created image sets using each of these filters, each with a variety of filtering thresholds (see Figure 2.5). We created degraded images at the following spatial frequency cutoffs: 4; 6; 8; 10; 12 Cycles/Face. We selected these frequency bands based on a pilot study which showed that they were the bands for which performance varied between the ceiling and floor accuracy for DCNN1 (see Appendix A).

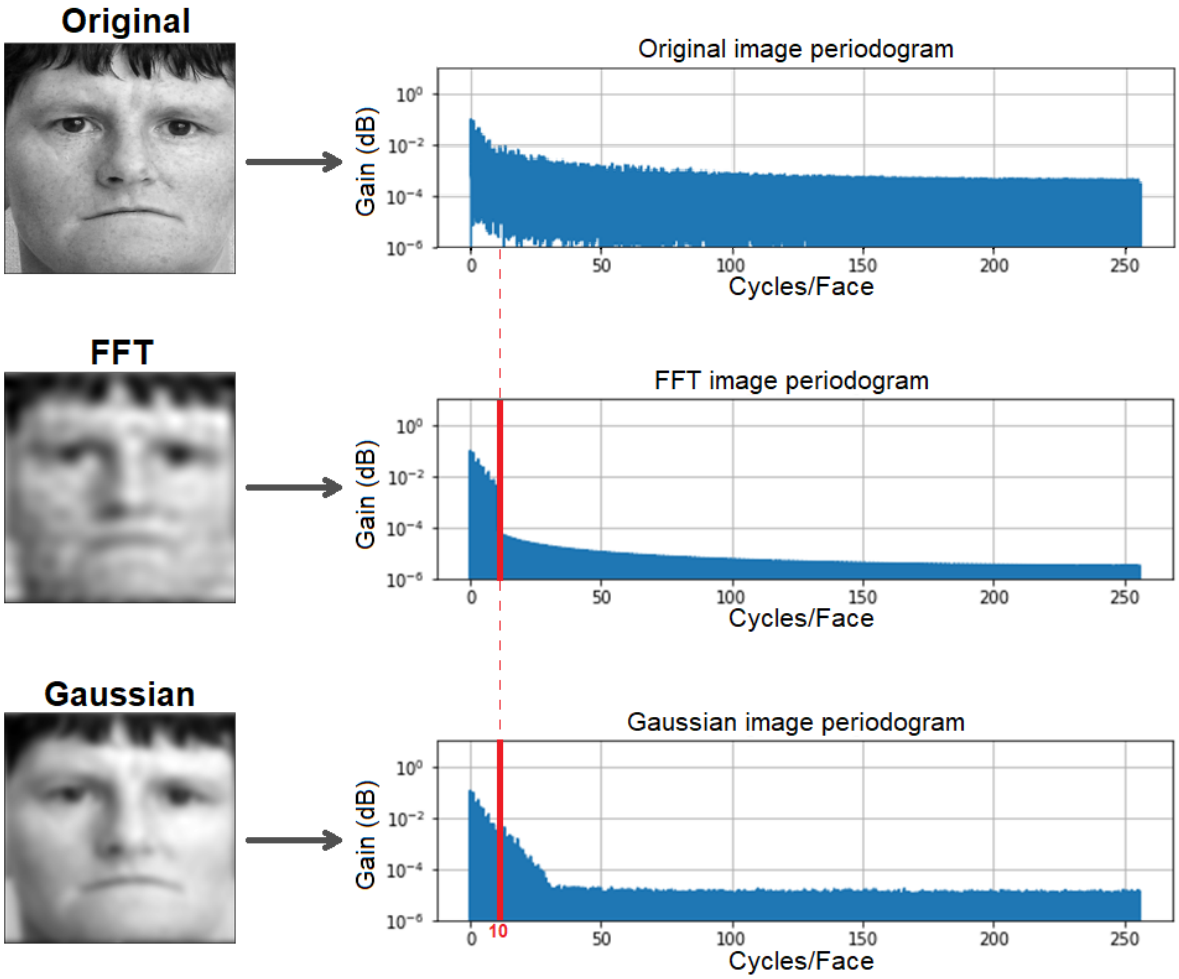


Figure 2.4. Example of low-pass filters on a 512x512 pixels image and periodograms containing spatial frequencies for each filter (set to 10 cycles/face). At the top, we show the original image. In the middle, we show the effect of the FFT low-pass filtering, which abruptly attenuates the information above the threshold, provoking the ringing effect (See Gibbs, 1898). At the bottom, we show the Gaussian low-pass filtering effect, which gradually attenuates the information above the threshold.

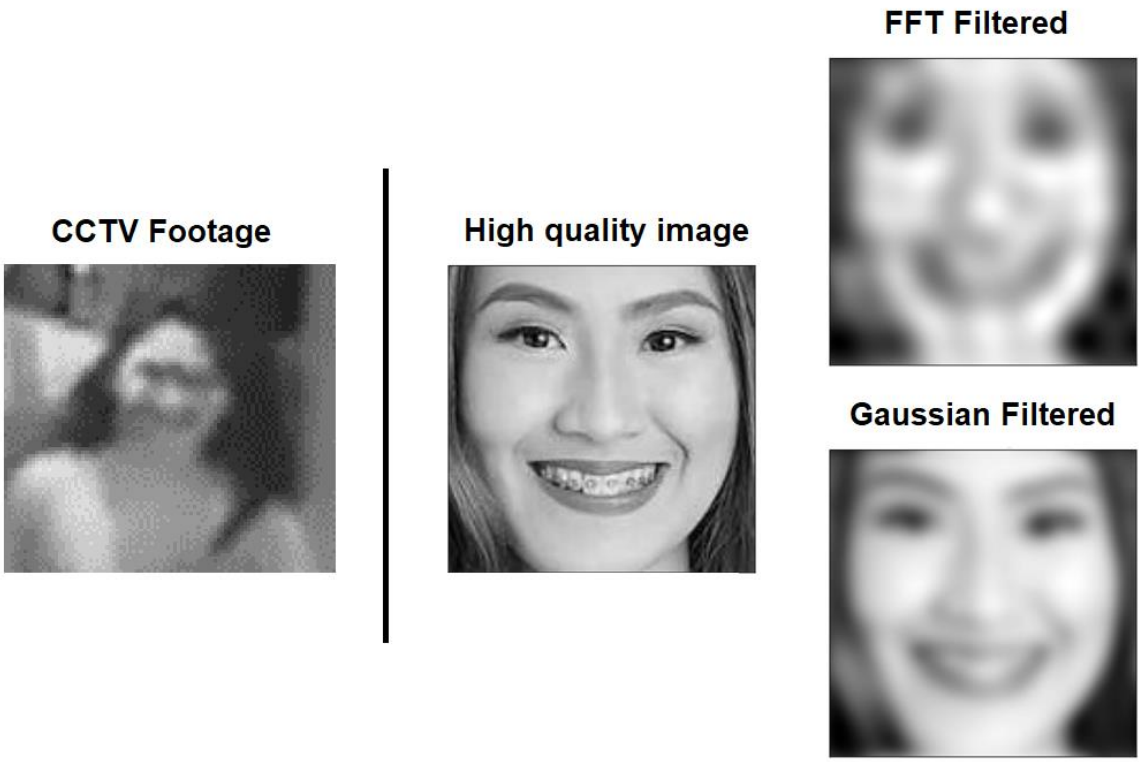


Figure 2.5. Simulation of distortions found in CCTV footage using FFT filtering. We show actual CCTV footage on the left, which contains distortions and low quality. On the right, we illustrate that FFT filtering could produce similar distortions when applied to a high-quality image of that identity compared to Gaussian filtering. We tuned the low-pass filters to 6 Cycles/face in this example. Photos from the case involving Christine Dacera (Wooley, 2021).

We next produced the matching task stimuli in two distinct presentation conditions, called Same-Resolution and Different-Resolution. For the Same-Resolution condition, the pair of face images making up a trial were either high-quality 'original' images or showed a similar level of filtering using the same filter. In contrast, for the Different-Resolution condition, one of the two images was always the original unfiltered, while the other showed a filtered image. Figure 2.6 shows examples of image pairs from all experimental conditions.

We prepared an identical set of stimuli for human participants and DCNNs, except that the images presented to DCNNs were 224x224 pixels and to human participants were 512x512 pixels (filtering levels were proportional). We changed the image size for humans because the DCNN images were too small to put on a computer screen side by side. So, we used the 512x512 to ensure that humans could see all the images at the traditional 60cm distance to the monitor, covering around 13º of visual angle per image. We presented all face images in grayscale.

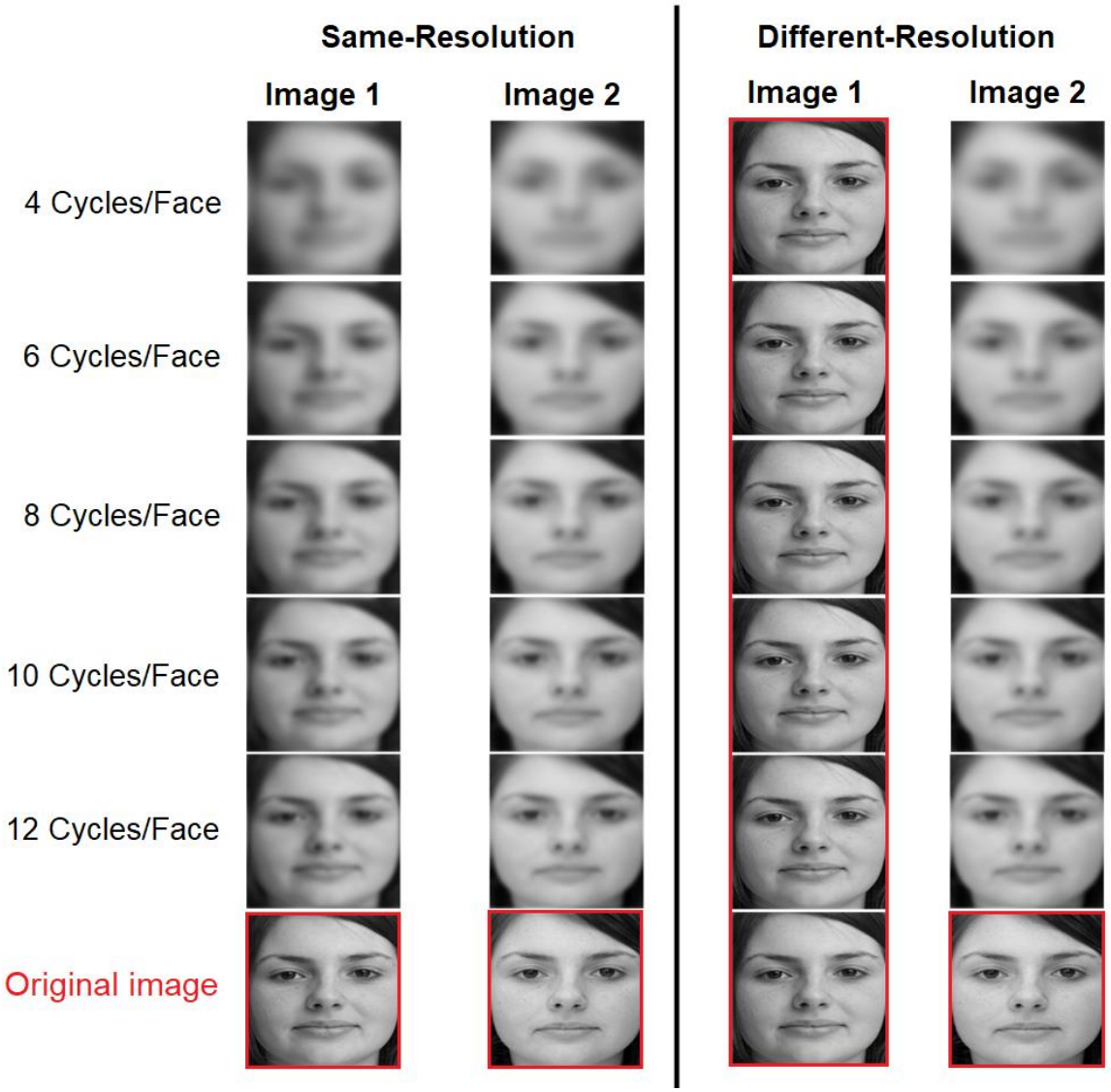


Figure 2.6. Examples of image manipulations on one of the face pairs used in this study. On stimuli Same-Resolution, we degraded both face images with a low-pass filter (FFT or Gaussian Blur). While on stimuli Different-Resolution, one image remained intact (Original image).

Procedure

Participants completed 50 trials (25 Match and 25 Non-Match), with each participant having a different random order of trials. We presented trials randomly to avoid similarity ratings being contextually derived immediate previous faces, aiming to reduce potential biases arising from serial dependence (see Liberman, Fischer & Whitney, 2014) and collecting more robust and unbiased data. We showed each trial in one of the six randomly selected spatial frequencies (4, 6, 8, 10, 12 cycles/face, and Original). Therefore, each participant saw a unique sequence of trials for the 50-item test. Also, It is important to note that because the allocation of trials to spatial frequency was random, some participants

might have seen only match (or non-match) trials at a particular spatial frequency, making it impossible to compute a given accuracy for that frequency. In these cases, we excluded that particular frequency for that participant (and related DCNNs), but this happened relatively rarely (i.e. ~2.5% of all frequency accuracies).

Each of the 9 DCNNs also processed the unique set of trials seen by each human participant. For the FFT versions, each DCNN processed 162 sequences of Same-Resolution trials and 168 Different-Resolution trials. For the Gaussian blur versions, each DCNN processed 100 same-resolution and 100 different-resolution sequences of trials. We did this to generate a distribution of DCNN scores at each spatial frequency on the test, comparable to that of the human participants.

Humans responded to the stimuli using a 5-point Likert scale (1: Sure Different, 2: Think Different, 3: Do Not Know, 4: Think Same, 5: Sure Same). We later enumerated these semantic values from 0 to 1 in steps of 0.25. Zero is equivalent to a 'Sure Different' response in these numerical responses, and one is equivalent to a "Sure Same" response. For DCNNs, we calculated similarity scores by inverse normalising the Euclidean Distances between two feature vectors. Therefore, for DCNNs, we would not see well-defined steps - as seen in humans - but a linear continuum ranging form 0 to 1. This continuum would represent the likelihood of the two identities being different (e.g. closer to 0) to the same (e.g. closer to 1). This normalisation process will still preserve the relative performance and discrimination thresholds, ensuring a fair comparison of models using features with different scales. And so, we used normalisation to enable visual inspection of human and algorithm data at the same range without affecting their individual outcomes (e.g. Phillips et al., 2018)

We calculated participants' overall accuracy using two distinct methods. First, we calculated each participant's accuracy on the full test across all stimulus conditions, using the area under the ROC curve (AUC). We call this analysis 'Non-Yoked by Cycles' (Non-Yoked) because it collapses the data across the different spatial frequency trials when computing AUC. In the second ‘Yoked by cycles’ method, we calculated AUC separately for each spatial frequency condition. Because the distribution of match and non-match similarity scores may vary in each of these conditions, these two methods were likely to produce different results. The reason we opted for the AUC (Area Under the Curve) as our method for calculating accuracy is because it has been used in human-machine performance comparison in prior work (e.g. Phillips et al., 2018; Phillips, 2017) and so provides continuity

with this work. It is also used as a standard way to assess accuracy when a continuous response scale is used and enables the accuracy of humans and machines to be directly compared, even where the response scales used by humans and machines may differ. Furthermore, the use of AUC to benchmark human and machine intelligence is not specific to the field of face identification but is common in other tasks (e.g. Jammal et al., 2020).

Results

This study had a factorial design with Participant Type (human and DCNNs) and Resolution (same and different resolutions) as between-subjects factors and Metric (yoked and non-yoked) as within-subjects factor. For the sake of clarity and brevity, only the results using the FFT filter are presented here. The results from the Gaussian filter - which produced similar results to the FFT – are presented in APPENDIX A.

Analysis of accuracy

First, we analysed data using a three-way ANOVA with three-way mixed ANOVA on the AUC data, with Participant Type (human and DCNNs) and Resolution (same and different resolutions) as between-subjects factors and Metric (yoked and non-yoked) as within-subjects factor. This analysis collapsed across the different levels of image resolution, although these are analysed separately in the following section. Figure 2.7 plots the accuracy results for the human participants and the average of the 9 DCNNs when attempting to match face pairs in the two conditions using the two methods used to measure accuracy. As described above, we employed two methods which we called Non-Yoked and Yoked by cycles. The Non-Yoked metric considers accuracy as a whole for the stimulus set. The Yoked calculates accuracy separately for each spatial frequency condition, and we calculate the overall accuracy score as the average of these different values.

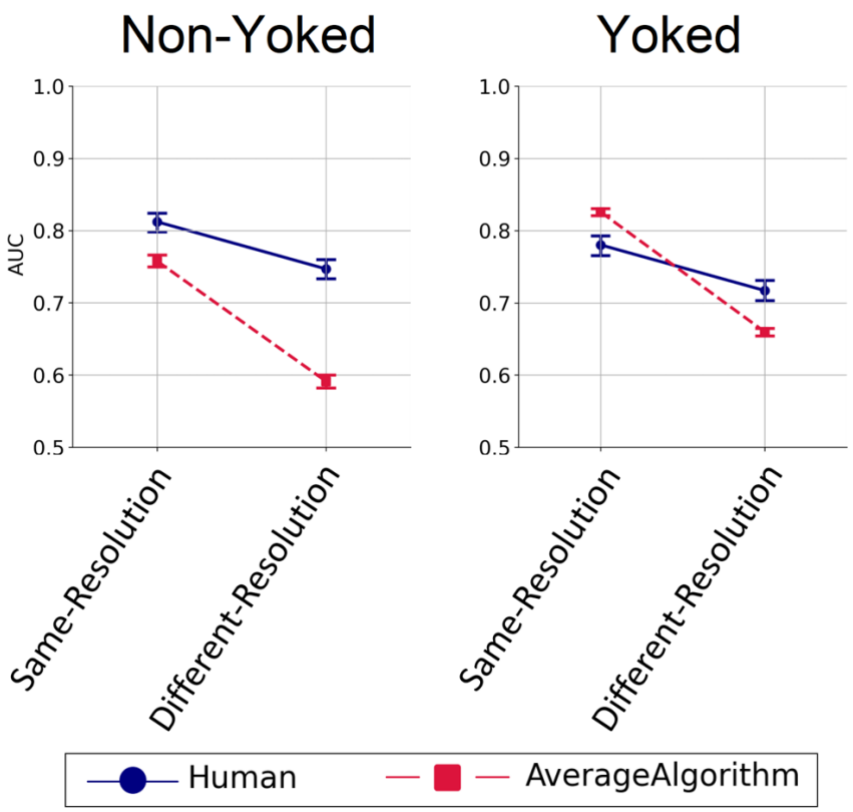


Figure 2.7. Graph showing the Resolution condition between Participant Type considering the two different Metrics to compute AUC. This graph shows the FFT version of the stimuli. The error bars show the 95% confidence interval.

We found that the main effect of Metric [F(1,1312)=22.73, p< 0.001, η^2p= 0.02] and the 2-way interaction between Participant Type and Metric were significant [F(1,1312)=160.51, p<0.001, η_p^2= 0.11], reflecting that human performance was more similar across the two measures of performance than was that of the DCNNs, where accuracy was higher in the Yoked condition. This difference is informative to the processing differences between DCNNs and Humans and is analysed further below (see 'Analysis of similarity ratings'). However, given the three-way interaction between Resolution, Participant Type, and Metric was not significant [F(1,1312)=0.01, p= 0.920, η^2p= 0.00], we next collapsed across the factor of Metric to examine the primary research question more closely – the differential effect of blurring on performance in humans and DCNNs.

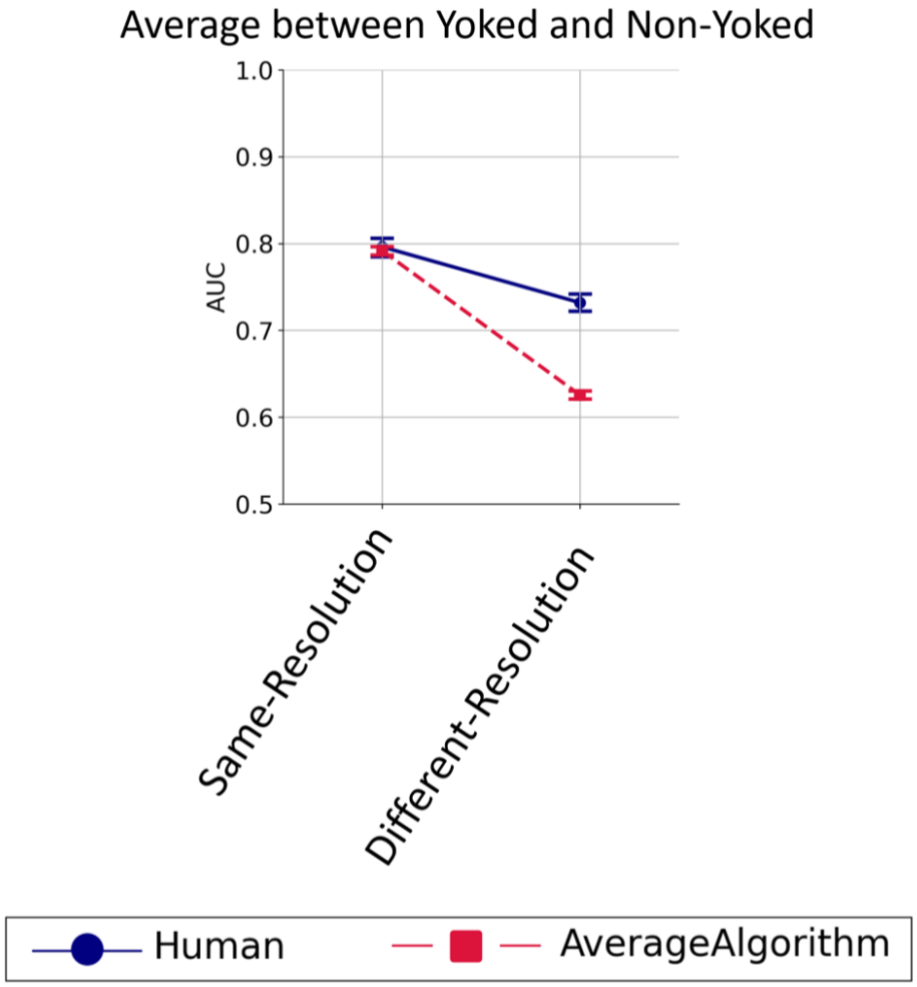


Figure 2.8. Graphs showing the effect of Resolution conditions for the FTT version of the study. Here, we calculated AUC as the average result between the Non-Yoked and Yoked metrics. The error bars show the 95% confidence interval.

We show the accuracy results of collapsing across the Yoked and Non-Yoked factors in Figure 2.8. We performed a two-way ANOVA on the average accuracy (AUC) data, with Experimental Conditions (same and different-resolution) and Participant Type (human and dcnns) as between-subjects factors. This analysis showed significant main effects of Experimental Condition [$F(1,656)=773$, $p< 0.001$, $\eta^2p= 0.541$], with higher accuracy overall for same-resolution image pairs, and Participant Type [$F(1,656)=179$, $p< 0.001$, $\eta^2p= 0.214$], with overall higher performance for humans. However, these main effects were qualified by a significant 2-way interaction [$F(1,656)=151$, $p< 0.001$, $\eta^2p= 0.19$], showing that DCNNs were poorer in the Different-Resolution compared to the same resolution condition [$t(656)= -28.36$, $p_{bonferroni}< 0.001$, Cohen's $d= -3.12$]. Humans were relatively less affected by the

change in resolution between comparison images [t(656)= -10.96, $p_{bonferroni}$< 0.001, Cohen's d= -1.2].

Analysis of similarity ratings

The ANOVA revealed that the impact of measuring performance using the yoked and non-yoked metrics was different for humans and DCNNs. For the DCNNs, the yoked metric gave higher performance estimates than the non-yoked metric, but the human performance was unaffected by the type of analysis. To examine the reason for this difference more closely, we examined the distributions of similarity scores generated by humans and DCNNs.

Figure 2.9 shows the Similarity-Scores distributions of human participants and one of the DCNNs (DCNN 1) for the data from the FFT filtered images. For clarity, we plotted only the data from DCNN 1 here because all DCNNs showed a similar pattern of results. The data from the Gaussian filtered version is available in APPENDIX A. Data in Figure 2.9 shows the similarity score distributions separately for the Same (top panel) and Different-Resolution (bottom panel) tasks. Each panel shows similarity scores across the six spatial frequencies for humans (left) and DCNN1 (right) and the Yoked and Non-Yoked accuracy metrics results.

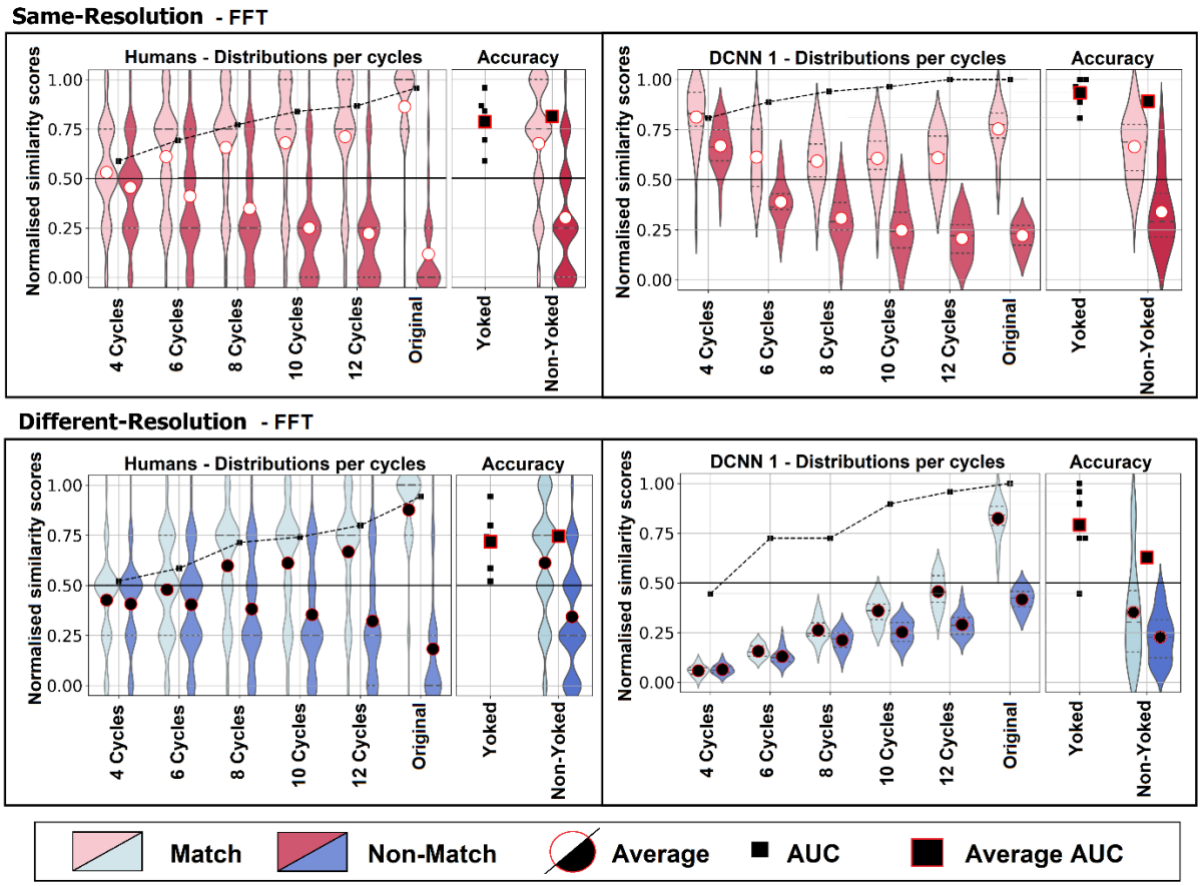


Figure 2.9. Distribution of Similarity-Scores across all participants and DCNN 1 in same (top row) and different-resolution (bottom row) conditions. The distributions on the left of each plot show match and non-match similarity scores for all spatial frequencies. The two rightmost distributions on each plot show accuracy (Area Under the ROC Curve – AUC), separately for the Yoked and Non-Yoked metrics. We show the Human data in the left plots and DCNN data on the right. An important difference between humans and DCNNs is that the human ratings tend towards the middle of the response scale as the images become of poorer quality (i.e. from right to left). For the DCNNs, poorer image quality causes ratings to tend towards one, or zero, depending on whether the image pairs were from the same image quality condition (top) or different image quality condition (bottom). The dotted lines connect the individual AUCs - plotted as a black square- calculated for each of the 6 spatial frequencies.

Inspection of Figure 2.9 shows that human participants (i.e. graphs on the left) were likely to use the middle point of the Likert scale ("Do Not Know", or 0.50) when image quality was poor. For DCNNs, on the other hand, poor image quality caused all similarity ratings to tend toward one (in the Same-resolution condition) and zero (in the Different-resolution condition) instead of adequately separating them into matching (i.e. towards one) and non-matching (i.e. towards zero) pairs. This fact explains why DCNNs accuracy was poorer in the Yoked versus Non-yoked analysis. Because DCNNs similarity ratings tend to

extremes for both match and non-match identity image pairs in more inferior resolutions, it has the effect of requiring a separate 'decision criterion' for each of the different image quality conditions. Human criteria were not affected by quality conditions because the same decision threshold could be used across conditions. That is, a single threshold near the scale's centre can achieve relatively optimal discrimination of same/different pairs across all the image quality conditions for humans due to the better calibration of match and non-match responding.

Analysis of feature similarity space in DCNNs

DCNNs describe each facial image as a numerical representation (feature vector), and we can investigate similarities between identities by comparing these descriptions. To explore the similarity space defined by these feature vectors, we used t-SNE (Maaten & Hinton, 2008) to visualise the similarity of feature vectors of DCNNs using every face present in the database (592 frontal faces, 296 identities) in each of the six filtering conditions (4, 6, 8, 10, 12 cycles/face, and Original). The t-SNE technique projects an n-dimensional similarity space defied by the n features in the penultimate DCNN layers onto a 2-dimensional space while maintaining local structure. This technique allows us to visualise the within and between identity variations in our image set.

We applied t-SNE (Maaten & Hinton, 2008) on DCNN 1 top-layer (2048 features) to visualise the similarity of every face in the database (592 frontal faces, 296 identities) in each of the six filtering conditions (4, 6, 8, 10, 12 cycles/face, and Original). As with the previous analysis, we show only the FFT filter dataset here, with the Gaussian filter dataset visualised in APPENDIX A.

We show the t-SNE similarity space in Figure 2.10. In an ideal face recognition algorithm, all images would be clustered by identity, with 296 identity clusters containing all 12 images (six image pairs) from the spatial frequency conditions. However, a visual inspection of Figure 2.10 shows a different pattern. While the images filtered at 8, 10, and 12 cycles are typically clustered by identity, original images - as well as the 4 and 6 cycles filtered images- generally are not found in these identity clusters. Instead, these images form their own clusters by image condition rather than by identity. This effect can be seen by the emergence of clusters of red (6 cycles) and blue (4 cycles) dots to the left of the plot

and black dots (original) to the right of the plot. This result helps explain why, for DCNNs, the same and different resolution conditions showed significant differences in performance. The same-resolution condition was easier for DCNNs compared to the Different-resolution condition because the similarity of image pairs was mostly driven by image quality over the identity information.

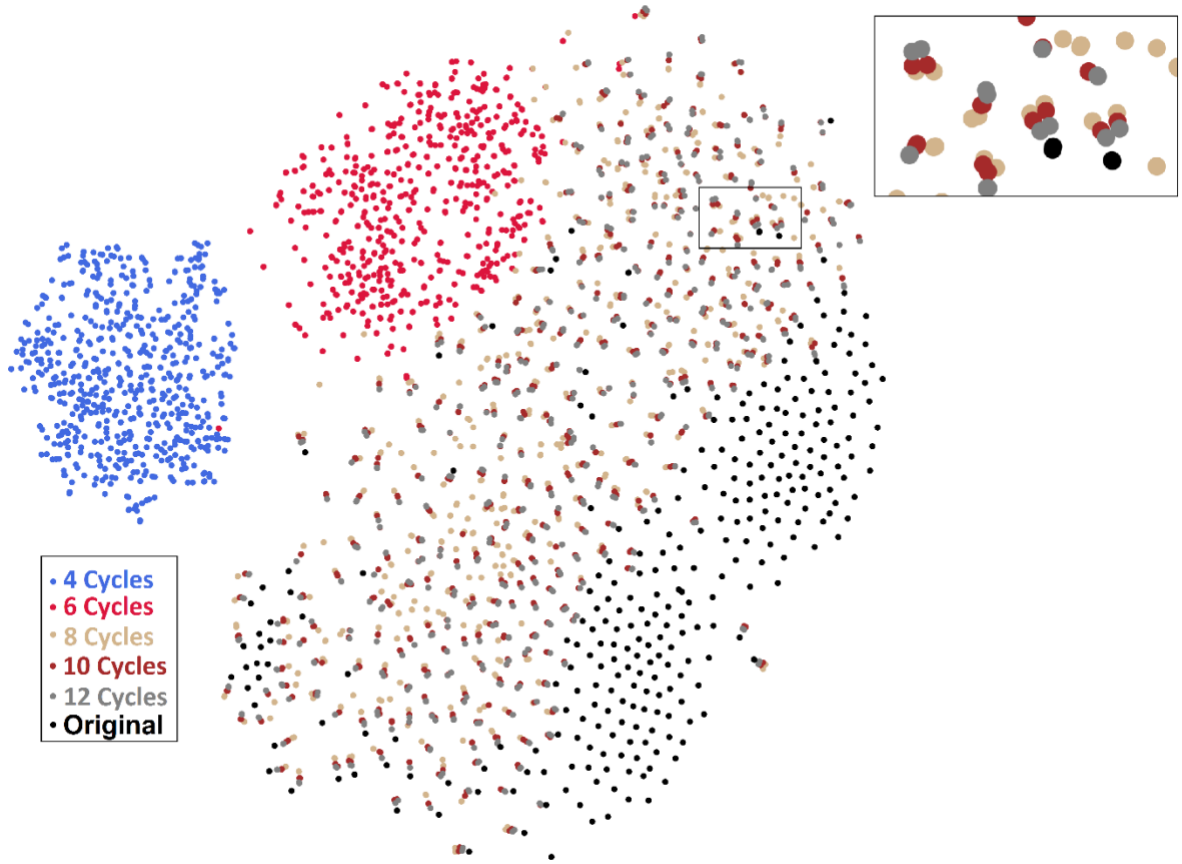


Figure 2.10. Two-dimensional visualisation of similarity space formed by 2048 features at the penultimate layer of DCNN 1 top-layer, with each dot being a face image from the database used in Experiment 1. In total, we split 592 face images of 296 identities by the six filtering conditions (4, 6, 8, 10, 12 cycles/face, and Original). We show examples of identity clusters in the 'zoomed in' region at the top left of the plot. Clear clusters of original, 4, and 6 cycles images show that similarity scores are often dominated by image quality rather than face identity. See "APPENDIX A – t-SNE visualisation" for t-SNE using Gaussian filter.

Discussion

In Experiment 1, we compared the accuracy of humans and previously trained DCNNs from the internet (see Table 1.1 for description) performing a face-matching test

under two conditions. We presented image pairs at the s*ame* or *different* resolutions. Our results show that the Different-Resolution condition was harder for both humans and DCNNs than the Same-Resolution one. However, we found that presenting the image pair at different resolutions had a greater impact on the performance of the DCNNs than it did for humans. This result is probably caused by better calibration of humans responding to the decision scale irrespective of image quality discrepancies, with the middle of the scale indicating uncertain judgements across all different levels of filtering. Thus, despite reducing image quality, humans used the response scale appropriately and consistently with matches above the midpoint and non-matches below the scale's midpoint.

In contrast, the DCNNs did not show this same degree of calibration. To compare the performance of DCNNs and humans more, we included the Yoked performance metric. The Yoked metric calculated accuracy separately for each level of filtering, thus, maximising results within each quality level. However, even when analysed this way, the accuracy of DCNNs is still degraded to a greater extent than humans when images are of different quality. We could argue that this greater decline in accuracy observed in DCNNs was because they have a more significant disruption in identity information when the images differ in image quality compared to humans. This result further suggests that the mechanisms used by humans and DCNNs are qualitatively different.

Research suggests a model in which humans process the high and low spatial frequency information in a face using two distinct cognitive channels (see Goffaux et al., 2005). Notably, the stimuli conditions in this study (i.e. Same and Different-Resolution) contained a mix of high and low-quality images. Humans may use these distinct cognitive channels to extract and process the available information to match the identities depicted in faces, allowing responses that are robust to changes in image quality. Stimuli trials in the Same-Resolution condition showed similar spatial frequency information. So, the model suggests that humans activated these cognitive channels similarly for the two faces in this condition. This similar activation of channels is because the amount of information in the two faces was almost identical. On the other hand, stimuli trials in the Different-Resolution condition showed distinct spatial frequency information between face pairs. Thus, the model suggests that humans had to extract only the comparable information between the two faces to decide whether the identity depicted in the two images was of the same individual or different people. In contrast, the lower accuracy of DCNNs when images were

of varying quality (i.e. in the Different-Resolution) may occur because DCNNs show significant impairment in extracting the usable 'configural' information from the high-quality images compared to humans. Interestingly, when we provided similar information (i.e. in the Same-Resolution), DCNNs showed comparable performance to humans in the FFT version of the stimuli (See APPENDIX A for the Gaussian version). This result illustrates that DCNNs can still perform the task, but humans demonstrated a qualitative superiority in extracting comparable information to process identities across pairs of images presented in different resolutions. See Figure 2.11.

Figure 2.11. Example of featural and configural information contained in a trial of the Different-Resolution condition. We observed significant reductions in accuracy for humans and DCNNs when processing trials of different than the same quality. Still, humans showed less impairment than DCNNs performing the task. We argue that the human advantage was because they could extract lower-quality configural information from high-quality images significantly better than DCNNs to process the stimuli. When showing images of comparable information (i.e. Same-Resolution condition), humans and DCNNs performed similarly.

Why are humans better able than DCNNs to extract information which is diagnostic of identity and stable across different resolutions? One possible explanation for this effect might be due to human development. Humans start to observe faces during the early stages of life and -importantly- with low visual acuity (see Rennels & Davis, 2008). This may provide

important experience in extracting identity-relevant information from low-quality images. On the other hand, DCNNS are trained primarily on high-quality images, and they do not have the same graded training experience as humans who are first exposed to low-quality images before seeing increasingly high-quality images as their visual system matures. Although it is not possible to provide the role of visual development in the results observed here, it is important to note that this is the first study to systematically compare humans and DCNNs processing faces of varying image quality. And so, limitations might arise when attempting such direct comparison, as the algorithms and humans operate with undoubtedly distinct training sets. As a result, it is still unclear whether the difference between humans and DCNNs is due to image quality differences or impairments in collecting valuable identity information within the face images. Unfortunately, addressing this matter comprehensively is beyond the scope of this thesis.

We used t-SNE (Maaten & Hinton, 2008) to visualise the structure in which DCNNs store identity information throughout the whole spatial frequency spectrum used in Experiment 1. The t-SNE plot presents similar DCNN descriptions of face images close to each other. However, the degradation of image quality caused the DCNNs to group images by a combination of identity and image resolution rather than by identity alone. We observe this effect in highly blurred images (e.g. 4 or 6 cycles/face). In these cases, the t-SNE compression shows that DCNNs perceived these highly filtered images as a distinct class of objects, detached from their original identity.

The t-SNE representation helps us understand why comparing facial images in different resolutions was particularly problematic for DCNNs – showing that the similarity of feature representations was often dominated by the effect of image quality, rather than the effect of face identity. This result is consistent with the findings of O'Toole et al. (2018), who depicted the structure of a DCNN space for a database of images scraped from the internet. Their results show low-quality images (due to poor resolution, occlusion, and blurring) located in a cluster in the centre of space. This result suggests that a significant factor determining the similarity of identities in DCNNs is the amount of quality signal contained in the image, which conflates with the identity signal itself.

To check that this pattern of results wasn't specific to the choice of filter adopted, we replicated the entire study using two different filters, FFT and Gaussian. To enhance readability, we have only presented the FFT data analysis in this chapter's main text. The

data from the Gaussian study are available in APPENDIX A. Inspection of that data reveals a very similar pattern of results to that reported here. Overall, the matching performance for the FFT-filtered images was less than the Gaussian-filtered images for both humans and DCNNs, but this may be because the FFT version added artifacts - such as "ringing" - in the image, which may have obstructed visual facial features. However, the pattern of results was similar for the two filters, suggesting that this may hold for other forms of image degradation, such as in authentic CCTV images.

Our results may have some practical significance. Although both humans and DCNNs were negatively affected by image quality degradations, humans outperformed DCNNs when comparing pairs of images of different quality. There are several essential settings where face comparisons often include images of varying quality—for example, comparing CCTV recordings against a reference face (e.g. passport images). Our results expand the practical understanding of using DCNNs in this type of scenario because we show that the sensitivity of DCNNs when attempting to verify identity information is somewhat proportional to the relative quality of its input images. However, we offer a solution to minimise the decrement in performance in DCNNs. In forensic settings, DCNNs compare facial images using a similarity score scale. And therefore, these systems probably use a particular position (e.g. a threshold) in this scale to determine the DCNN's agreement regarding an individual's identity. Here, by showing that the Yoked metric led to improvements in decisions, we argue that a possible solution to minimise errors is to define multiple thresholds to make facial verification decisions. Notably, such thresholds would change depending on the quality of the images.

It is essential to address that we might have overestimated the power of the DCNNs in this study. The databases used to train the DCNNs employed in this study mostly contained high-quality images (e.g. VggFace2 (Cao et al., 2018), CASIA-WebFace (Yi, Lei, Liao, & Li, 2014), etc.). So, the performance of the DCNNs when matching lower-quality images might have been improved if we first used techniques to enhance image quality. Furthermore, it is unclear how the DCNNs would perform if they were trained on a more varied image set or if training mimicked the developmental processes in childhood, where image quality was initially low but improved over time (Dobson, Teller, & Belgum, 1978; Vogelsang et al., 2018; Jang & Tong, 2021;). Future work should expand our comparison analysis between humans and DCNNs by once improving the quality of images before the

examination (see for review Liu, Pedersen, & Wang, 2022) and by using promising DCNNs that comprehend lower-quality images (see Zangeneh, Rahmati, & Mohsenzadeh, 2020). In addition, one aspect that warrants closer examination is the understanding of the use of Likert-scales and Euclidean Distance by both humans and DCNNs - as our full understanding of how they represent responses remains incomplete. And so, future studies should focus on elucidating how similar these scales are and how they contribute to the similarities observed between the two distinct entities.

Nevertheless, we show decrements in image quality negatively impact both humans and current DCNNs. In forensic settings, humans often employ DCNNs to help guide their decision-making. And despite humans and DCNNs being error-prone in performing facial identification, previous research has also shown an exciting solution to improve false allegations in forensic settings. This body of research shows that combining (i.e. fusing) decisions made by the best human operators alongside DCNNs matching identities in faces could improve accuracy to reach ceiling levels (e.g., Phillips et al., 2018; White, Dunn, Schmid & Kemp, 2015). Indeed, our results outline the potential for fusing decisions made by humans and DCNNs. Previous research shows that groups of individuals with different strategies for solving problems are better than a single source (Kittler, Hatef, Duin & Matas, 1998; Hong & Page, 2004; O'Toole, Abdi, Jiang & Phillips, 2007; White, Burton, Kemp & Jenkins, 2013; Hu et al., 2017; Jeckeln et al., 2018). And so, it is plausible to interpret that grouping human responses with DCNNs' would theoretically benefit the overall accuracy due to qualitative differences between their judgment about identities in faces. However, it is still unclear how to properly use the information provided by humans and DCNNs to use them best. Thus, the next chapter of this thesis will focus on finding the best practices to improve decisions made by simulating such forensic "teaming" between humans and DCNNs.

Conclusion

Here, the stimulus presented reinforces that there are important differences in the performance of humans and algorithms when making face-matching decisions. Humans showed substantial advantages relative to DCNNs in processing image pairs that differed in quality. As a result, this work raises a new argument showing that humans share distinct but

complementary processes (i.e. different strengths) compared to algorithms processing faces of diverse image qualities.

Chapter 3 - Optimising human and AI teaming by being sensitive to individual differences

Introduction

In many applied settings, it is necessary to establish the identity of an unknown person, such as when identifying travellers at border crossings and establishing the identity of a fugitive in a criminal investigation. These important identification decisions must be made with extremely high accuracy because errors can seriously affect people's lives and public safety. Organisations will typically have human staff make these face identification decisions with the assistance of face recognition technology (built using artificial intelligence, e.g. DCNNs). This is commonly referred to as "human-AI teaming", or "human-AI fusion".

The work of Phillips and colleagues (2018) showed that combining facial identification decisions made by humans and DCNNs can improve the accuracy of the decisions up to ceiling levels. In a face-matching task where two images were shown simultaneously on a screen, human experts (facial forensic examiners) made identity judgments using a 7-point Likert scale (-3 = Very confident different people; +3 = Very confident same people). When averaging these ratings with similarity scores produced by leading DCNNs (~95% accuracy) and a forensic examiner (~93% accuracy on average), the combination resulted in facial identification decisions reaching 100% (median of distribution). And so, their work provides a practical and significant method to illustrate that humans and DCNNs can be powerful tools for making facial identification decisions when working together.

However, such a human-AI combination can also decrease the quality of decisions. That is, they also show that using the same forensic examiners combined with DCNNs of poorer quality (~67% accuracy) can reduce the overall decisions to 91%. Thus, forensic examiners alone would be a better choice in this case - as their accuracy alone was higher than the human-AI. Although both humans and DCNNs are capable of high performance, they still make errors. There is, therefore, a practical need to devise methods that optimise how humans and AI work together to identify faces.

One potential option to improve the accuracy of human-AI teaming is to improve the accuracy of the humans. In typical participant samples, the distributed range of measured face processing ability goes from individuals with an extremely poor ability (i.e. individuals diagnosed with developmental – or acquired - prosopagnosia (for reviews, see Bate &

Bennetts, 2014; DeGutis, Cohan, & Nakayama, 2014), to individuals on the other side of the spectrum (i.e. super-recognisers (see Russell, Duchaine & Nakayama, 2009; Bobak et al., 2016; Ramon, Bobak & White, 2019). In addition, research suggests that ability is highly stable over time (Sutherland et al., 2020; Germine et al., 2015), with test re-test correlations above 0.7 (Balsdon et al., 2018; White et al., 2021), and highly heritable (Wilmer et al., 2010, Zhu et al., 2010; Shakeshaft & Plomin, 2015). And so, a possible approach to improving human accuracy would be to select individuals on the basis of this skill, but there is no principled basis on which to select individuals that are sufficiently skilled to form an AI-human team.

Alternative approaches include training individuals to maximise their identification performance. However, these attempts were rarely successful for individuals diagnosed with prosopagnosia (DeGutis, Cohan & Nakayama, 2014; DeGutis et al., 2013; Ellis & Young, 1988; Brunsdon, Coltheart, Nickels, & Joy, 2006; see Towler et al., 2021 for a review). While some simple approaches, such as providing accuracy feedback (White et al., 2014), might slightly improve someone's facial processing, this improvement is only found for participants initially showing poorer skills. Despite numerous training courses developed to improve identity verification abilities (see Moreton, Havard, Strathie, & Pike, 2021), a comprehensive evaluation of their outcomes showed they could not significantly improve face-matching accuracy (Towler et al., 2019).

Another potential solution to improve the accuracy of human-AI teaming is to remove the human from the decision so that the algorithm makes identity verifications alone. Current DCNN technology possesses a robust designed architecture that can recognise thousands of learned identities with outstanding performance (e.g. Cao et al., 2018; He, Zhang, Ren & Sun, 2016; Schroff, Kalenichenko, & Philbin, 2015). However, today, despite DCNNs being accurate and significantly fast (i.e. seconds to make a final decision), their accuracy in performing unfamiliar identity verifications is comparable to super-recognisers' (Grother et al., 2019).

But entrusting algorithms completely with face identification decisions is problematic. For example, as shown in the previous Chapter 2 of this thesis, algorithms are susceptible to image quality issues, leading to potential errors. It is perhaps not surprising then that when algorithm accuracy has been evaluated 'in the wild' – i.e. in operational settings - official reports find poor facial recognition performance or many false positives.

For example: (i) the Welsh police tried implementing algorithms to compare the identities of 170,000 people against a database of custody images during a football match. During this implementation, the algorithm found 2,470 possible matches, but 92% of those identities (2,297) were false positives (Press Association, 2018); (ii) In the United States, an algorithm used by police departments compared 28 images of members of Congress against a database of 25,000 mugshots. In this operation, the algorithm detected 28 possible matches (5% error rate). Interestingly, most of these misidentifications were from Latino and African-American members (Singer, 2018); (iii) also in the United States, an algorithm implemented to compare surveillance CCTV footage against a database of ID photographs mistakenly identified a man as a criminal. Such an accusation - being trusted by the police - led this man to be imprisoned for nearly 30 hours (Williams, 2020). Such examples demonstrate that facial recognition technology cannot be used for facial verification without some level of human oversight.

Human oversight of automatic facial identification decisions is not always effective. In security settings, an important use of face recognition software is to allow a larger volume of identity verifications, given their speed and accuracy. Such a service allows these settings to use face recognition software to quickly scan a facial probe image and compare its identity against vast face databases to find similar-looking individuals. This search process is known as one-to-many (i.e. 1-to-n), and its output is a list of similar-looking candidates (Jain, Klare, & Park, 2012; Grother, Ngan, Hanaoka, 2014; Jain & Ross, 2015). However, as White and colleagues (2015) demonstrated, using such auxiliary tools to find possible candidates is problematic as humans are still required to make the final decision regarding the target's identity. In their study, they compared groups of untrained and trained individuals when asked to find a target identity amongst an algorithm's list of eight potential candidates. They reported that trained staff members (i.e. passport issuance staff, who make an average of 60 facial identification decisions using face recognition software per day) were no better than untrained individuals when locating targets. And the untrained group made over 50% of target identification errors. Interestingly, they also tested a select group of specialist facial examiners performing the same task and found that they performed 20% better than the other groups (White, Dunn, Schmid, & Kemp, 2015). So, adding humans to oversee DCNNs is arguably ineffective in reducing error rates in facial identifications.

In addition, when the task is to compare identity pairs (i.e. a 1-to-1 verification), studies show that when humans have access to the algorithm's decisions regarding their similarities, humans have a significant tendency to bypass its decision. That is, not capturing errors made by the algorithm (e.g. Fysh & Bindemann, 2018; Howard, Rabbitt, & Sirotin, 2020). For example, Howard and colleagues (2020) created an expanded version of the GFMT (Burton et al., 2010). For processing such a task, participants should decide the similarities of identity pairs - sequentially presented on a screen – by ranking their identity similarities on a 7-point Likert scale. However, in some conditions of their task, they added a background colour to the stimuli, indicating whether an 'algorithm' - or a human - previously decided that such an identity pair was from different or the same identities. Their results show a significant shift in human responses to accept the previous decisions, being them from humans or algorithms. Thus, such results further illustrate that a potential sequential human oversight regarding AI's facial identification decisions will likely not improve the quality of decisions.

Where human-AI teaming works well appear when they perform the task independently but equally contribute to an overall decision. Laboratory studies show that the 'wisdom-of-crowds' significantly improve face-matching accuracy. For example, when humans perform independent decisions comparing the identities of individuals (e.g. on a 5-point Likert scale), studies show that the optimal outcome is likely to be found in the average of multiple decisions (Dowsett & Burton, 2015; Jeckeln et al., 2018). Such a method for aggregating responses is called 'wisdom-of-crowd'. Ultimately, when there is more diversity in the strategies employed by the decision-makers, averaging their decisions (e.g. fusing responses) results in even more accurate predictions to match identities (see Kittler, Hatef, Duin & Matas, 1998; White, Burton, Kemp & Jenkins, 2013; Hu et al., 2017; Jeckeln et al., 2018). Interestingly, applying the same concept of diversified independent sources by pairing humans and AI can significantly improve the quality of the identity verification decisions (see Phillips et al., 2018, Towler et al., under review), producing better outcomes when compared to the two individual decision-makers alone.

The work of Phillips and colleagues (2018) provides evidence regarding humans-AI fusion improving face identification decisions. However, because they did not investigate the relations between the decisions made by humans and DCNNs, it raises questions regarding the underpinnings of such improvements. For example: Can individual differences

explain the size of such fusion effects? Or even, will decision differences between humans and DCNNs improve fusion effects even more? These questions are a critical implication of their study because unfamiliar face recognition is responsible for large individual differences in humans. And so, one possible route for studying the behavioural outcomes of fusion effects is to study its interaction with individuals situated in different places of the ability spectrum. Such a study will not only allow us to clarify how the face-processing ability of individuals would interact with DCNNs but also help us to find best practices to improve - and possibly predict - the quality of forensic identifications.

To our knowledge, no research has investigated the relationship between the face recognition abilities of individual humans and DCNNs - and how much this impacts the benefit of fusing their face identity processing decisions. In addition to showing that face processing ability has significant individual differences (Duchaine & Nakayama, 2006; Burton, White & McNeil, 2010; Phillips et al., 2018; for review, see White & Burton, 2022), the accuracy of individual DCNN algorithms also vary substantially from one algorithm to the next (Phillips et al., 2018; Towler et al., under review; Grother, Ngan & Hanaoka, 2019). And so, it is important to know how the relative levels of accuracy of both DCNNs and humans affect their combined accuracy. Therefore, we designed the study presented in Chapter 3 to understand better how individual differences in human face-matching accuracy can affect the size of fusion effects. This study aims to optimise processes for combining humans and AI to improve face-matching decisions in forensic scenarios.

Computational study 1

This study aims to better understand how humans and DCNNs can best complement each other to improve the quality of face-matching decisions. We used the data obtained from Experiment 1, where both humans and DCNNs showed variable accuracy on the task. Because perfect accuracy was rare, given the varying image quality of the dataset, fusing the decisions of humans and DCNNs from this task can show the degree of improvement when performing a difficult test. Because there were large variations in accuracy across individual humans and individual DCNNs, this also provides a suitable dataset for examining how best to combine individual humans with individual algorithms in an AI-human team.

Method

Participants

Participants were the same as in Experiment 1, where 513 student participants' performed the Same and Different-Resolution conditions in the Fast Fourier Transformation (FFT) and Gaussian Blur versions (see Chapter 2).

Stimuli

We used the same stimuli described in Experiment 1 (see Chapter 2). As a reminder, we produced the matching task stimuli in two distinct presentation conditions: Same-Resolution and Different-Resolution. In the Same-Resolution condition, participants and DCNNs processed 50 face pairs. We degraded the two face images of each trial using a low-pass filter set to 6 different randomly assigned cut-off frequencies. In the Different-Resolution condition, one facial image remained original while the filter degraded the other. We had two versions of these stimuli, one using an FFT filter and the other using a Gaussian Blur filter. We used both the FFT and Gaussian versions as stimuli for this investigation. However, their outcomes were analysed separately to keep consistency regarding the previous experiment.

Procedure

In Experiment 1, human participants responded to 50 face-matching trials using a 5-point Likert scale (1: Sure Different, 2: Think Different, 3: Do Not Know, 4: Think Same, 5: Sure Same). To analyse the data, we rescaled these responses from 0 (Sure Different) to 1 (Sure Same). For the DCNNs, we inverse-normalised the Euclidean Distance between the penultimate (top) layers (feature vectors) for each stimuli pair. As a result, both human and DCNNs' similarity scores were in the same range from 0 to 1, where 0 is equivalent to a 'Sure Different' response and 1 a 'Sure Same' response. Importantly, humans and DCNNs processed trials independently. This way, humans could not be influenced by DCNNs during the experiment.

Because the similarity scores of humans and DCNNs were in the same range (i.e. scale), we could combine them into one "fused" similarity score. Phillips and colleagues (2018) suggest fusing the responses of humans and DCNNs by averaging trial responses. And so, we replicated this same analysis in this study.

Importantly, in the previous Chapter 2, we found that DCNNs had substantial differences in the distribution of similarity scores for different image qualities compared to humans. Humans tended to calibrate their responses similarly across image quality conditions because they centred their overall score on the scale's midpoint, but this was not the case for DCNNs. DCNNs showed very different response distributions across the different image quality conditions. Therefore, to address this difference between humans and DCNNs in the current study, we processed DCNNs' similarity scores using two separate methods: (i) normalising all similarity scores together, combining all image quality conditions (or 'Direct Fusion'); (ii) normalising separately for each image quality band (or 'Quality-sensitive fusion'). Normalising separately for each image quality condition would likely improve fusion scores because it would force the average of all quality conditions to be similar on the scale. However, although this approach may not always be practicable in applied settings as it would require an extra step to rank the quality of an image, we present both types in our analysis. Figure 3.1 illustrates these different fusion approaches.

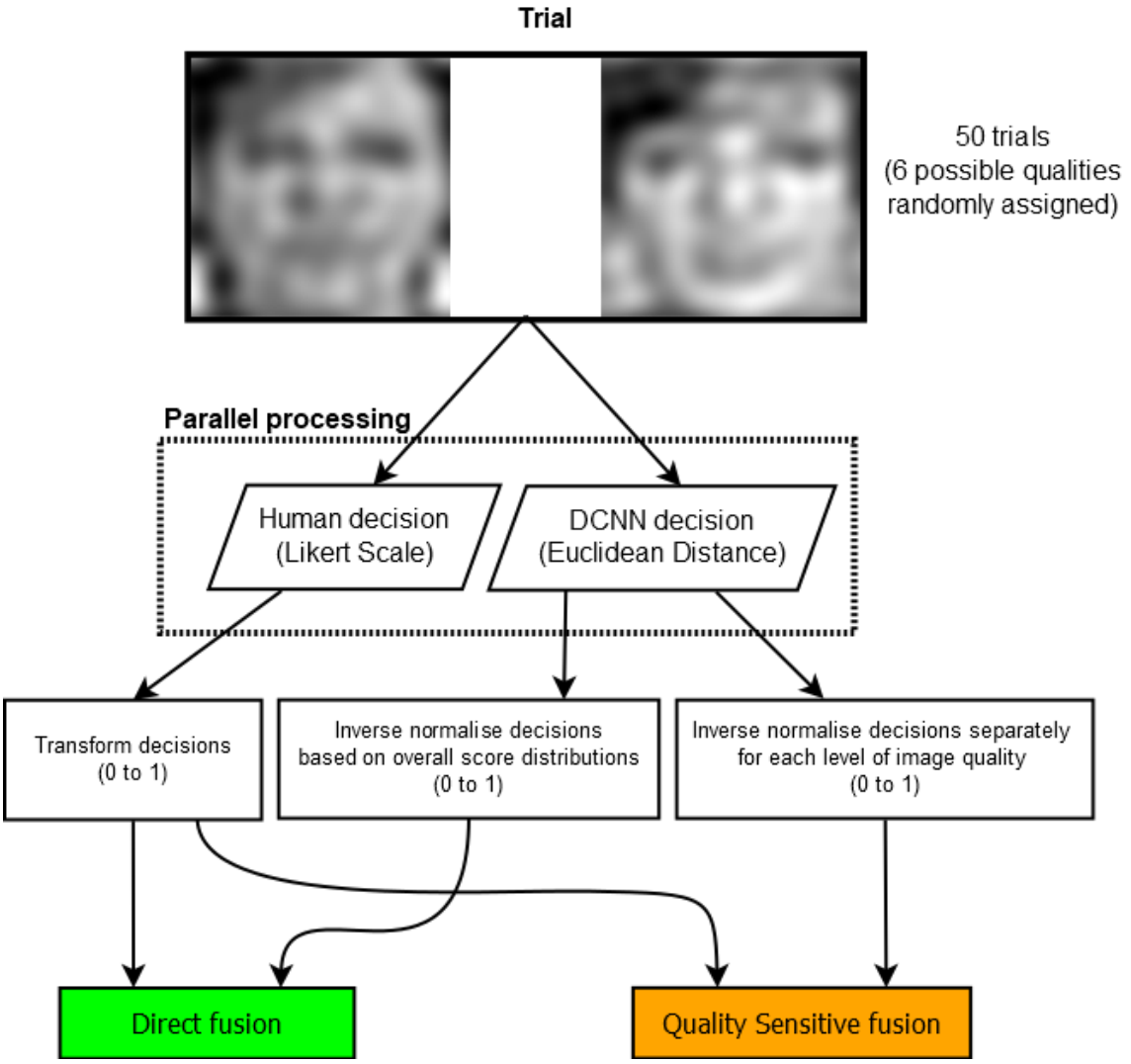


Figure 3.1. Schematics for the two metrics measuring fusing decisions made by humans and DCNNs processing faces. Humans and DCNNs processed 50 face-matching trials independently. Later on, we transformed the responses of humans and DCNNs for them to be in the same range (from 0 to 1). The fusion is the average trial response from humans and a DCNN.

After normalisation, we fused the decisions of individual DCNNs with individual humans. Our objective with the fusion was to investigate how the decisions of DCNNs and humans combine and improve the quality of the decision regarding someone's identity. And so, to fuse their decisions, we collected the similarity scores made by individual humans and a DCNN performing the same stimuli and averaged their responses. Here, we fused human responses with all nine individual DCNNs in the study (see Chapter 2 for the description of DCNNs). In addition, we created a tenth DCNN, made by combining (i.e. fusing) all the nine individual DCNNs together. We called this fused DCNN 'Average DCNN'. The fusing process of humans and DCNNs created a new set of 10 'fused' responses. We used these new

responses to verify if the fusion improved the decision quality compared to the ones made by humans or DCNNs alone.

We predict that fusing humans and DCNNs will produce more accurate decisions compared to individual decisions. To make this comparison, we calculated AUCs using the similarity scores from each of the three decision-makers (i.e. humans, DCNNs, fusion) as a measure of decision quality. And so, the difference between these computed AUCs will allow us to investigate the effects of fusion in the human-DCNN teaming for the 10 DCNNs. We replicated this process using the Direct and Quality Sensitive fusion metrics (see Figure 3.1).

Because we predict that the differences in how humans and DCNNs rank similarities will also lead to differences in their fusion outcomes, we measured their agreement between the decisions. For instance, studies show that diversity in the strategies employed by the decision-makers would improve the overall quality of the decision when matching faces (e.g. White, Burton, Kemp & Jenkins, 2013). And so, to investigate the monotonic agreement between humans and DCNNs, we calculated the Spearman's rho between their similarity scores as the agreement measure. Investigating how the agreement between humans and DCNNs can interact with fusion improvements is of great importance because it might give us an additional tool to predict the effects of this fusion posteriorly.

To improve readability, we present the detailed data analysis using the FFT version of the stimuli below. For analysis using Gaussian filter - which produced similar results to the FFT - see APPENDIX B.

Results

Human-DCNN fusion improvements correlate with human performance

We first verified if fusing human decisions with DCNNs could boost accuracy relative to the DCNN accuracy alone. For that, we calculated the Spearman's rho (**ρ**) between the AUC boost caused by the fusion ($AUC_{Boost} = AUC_{Fusion} - AUC_{DCNN}$) and the accuracy of the individual human that was used in the fusion (AUC_{Human}). We calculated these correlations for the two stimuli conditions (Same-Resolution and Different-Resolution) using the two fusion methods (Direct fusion and Quality Sensitive fusion). Here, we report our results in terms of the fusion between humans and the Average DCNN mainly to improve readability,

as its consolidated results suffice to represent the consistent findings across all DCNNs (for description, see Table 1.1, page 41 of this thesis document). Still, when necessary, we show a systematic comparison across each individual DCNN for comprehensive overview of the whole data. We found that for the Same-Resolution condition, both the Direct [**ρ**(160)= 0.56, p< 0.001] and Quality Sensitive [**ρ**(160)= 0.79, p< 0.001] fusion methods showed significant correlations with human performance. We found a similar pattern for the Different-Resolution condition. That is, both the Direct [**ρ**(166)= 0.74, p< 0.001] and Quality Sensitive [**ρ**(166)= 0.79, p< 0.001] fusion methods showed significant correlations with human performance. Moreover, all the other 9 DCNNs in this study showed a similar pattern of results when replicating the same analysis, which shows that the better individual human performers are, the greater the boost they can provide to DCNN performance after fusion (see Table 3.1).

	Same-Resolution		Different-Resolution	
	Direct Sensitive	Quality Sensitive	Direct Sensitive	Quality Sensitive
DCNN1	**ρ**(160)= 0.64, p< 0.001	**ρ**(160)= 0.79, p< 0.001	**ρ**(166)= 0.72, p< 0.001	**ρ**(166)= 0.82, p< 0.001
DCNN2	**ρ**(160)= 0.74, p< 0.001	**ρ**(160)= 0.71, p< 0.001	**ρ**(166)= 0.75, p< 0.001	**ρ**(166)= 0.75, p< 0.001
DCNN3	**ρ**(160)= 0.55, p< 0.001	**ρ**(160)= 0.72, p< 0.001	**ρ**(166)= 0.76, p< 0.001	**ρ**(166)= 0.77, p< 0.001
DCNN4	**ρ**(160)= 0.54, p< 0.001	**ρ**(160)= 0.70, p< 0.001	**ρ**(166)= 0.74, p< 0.001	**ρ**(166)= 0.83, p< 0.001
DCNN5	**ρ**(160)= 0.54, p< 0.001	**ρ**(160)= 0.67, p< 0.001	**ρ**(166)= 0.74, p< 0.001	**ρ**(166)= 0.75, p< 0.001
DCNN6	**ρ**(160)= 0.55, p< 0.001	**ρ**(160)= 0.74, p< 0.001	**ρ**(166)= 0.80, p< 0.001	**ρ**(166)= 0.83, p< 0.001
DCNN7	**ρ**(160)= 0.63, p< 0.001	**ρ**(160)= 0.79, p< 0.001	**ρ**(166)= 0.81, p< 0.001	**ρ**(166)= 0.84, p< 0.001
DCNN8	**ρ**(160)= 0.76, p< 0.001	**ρ**(160)= 0.77, p< 0.001	**ρ**(166)= 0.80, p< 0.001	**ρ**(166)= 0.82, p< 0.001
DCNN9	**ρ**(160)= 0.67, p< 0.001	**ρ**(160)= 0.71, p< 0.001	**ρ**(166)= 0.81, p< 0.001	**ρ**(166)= 0.85, p< 0.001
Average	**ρ**(160)= 0.56, p< 0.001	**ρ**(160)= 0.79, p< 0.001	**ρ**(166)= 0.74, p< 0.001	**ρ**(166)= 0.79, p< 0.001

Table 3.1. Table showing Spearman's correlations between AUC boost caused by the fusion ($AUC_{Boost} = AUC_{Fusion} – AUC_{DCNN}$) and the accuracy of the individual human that was used in the fusion (AUC_{Human}). In Table 3.1, we show these correlations for the two stimuli conditions (Same-Resolution and Different-Resolution) using the two fusion methods (Direct fusion and Quality Sensitive fusion).

In addition, we noted that the quantity of humans that could boost the DCNN accuracy ($AUC_{Boost} > 0$) relates to the DCNN used. For example, using the Direct fusion method in the Same-Resolution condition, we found that ~89% of our human participants could positively boost the Average DCNN's performance. However, a systematic view looking at individual DCNNs showed substantial variation in this figure, for example, only ~37% of the participants boosted DCNN2's performance. This effect is likely due to the fact that some DCNNs performed the task better than others (e.g. Average DCNN: M=80.28%, SD= 6.89%; DCNN2: M=93.14%, SD= 2.29%; t(160)=21.84, p<0.001).

Our finding that human performance can boost DCNNs' decisions has practical significance because it shows a valuable way to use humans and DCNNs to achieve optimal performance (see Chapter 2). In this initial investigation, we observed a linear relationship between the fusion effect and the face-processing ability of humans. Because human ability is stable across the whole performance spectrum (e.g. see White et al., 2021), this may point to a principled way of selecting humans to form an AI-human 'team' given a particular DCNN in question. We examine this possibility in the next section using a more formal analysis.

Human-DCNN fusion is improved when human performance is within 10% of DCNNs' performance

We aimed to understand how best to select humans for AI-human teams given a particular level of discrepancy between DCNN and human accuracy. To do this, we calculated the Spearman's rho (ρ) between the AUC boost caused by the fusion ($AUC_{Boost} = AUC_{Fusion} - AUC_{DCNN}$) and the difference between human-DCNN performance ($AUC_{Difference} = AUC_{Human} - AUC_{DCNN}$). Figure 3.2 shows scatterplots of these correlations for the two fusion methods (left: Direct fusion; right: Quality Sensitive fusion) for both the Same-Resolution condition (top panel) and Different-Resolution condition (bottom panel). For the Same-Resolution condition, both the Direct [$\rho(160) = 0.92$, $p < 0.001$] and Quality Sensitive [$\rho(160) = 0.91$, $p < 0.001$] fusion methods showed significant positive correlations with the performance difference between humans and DCNNs. The Different-Resolution condition showed similar high correlations, both the Direct [$\rho(166) = 0.92$, $p < 0.001$] and Quality Sensitive [$\rho(166) = 0.95$, $p < 0.001$].

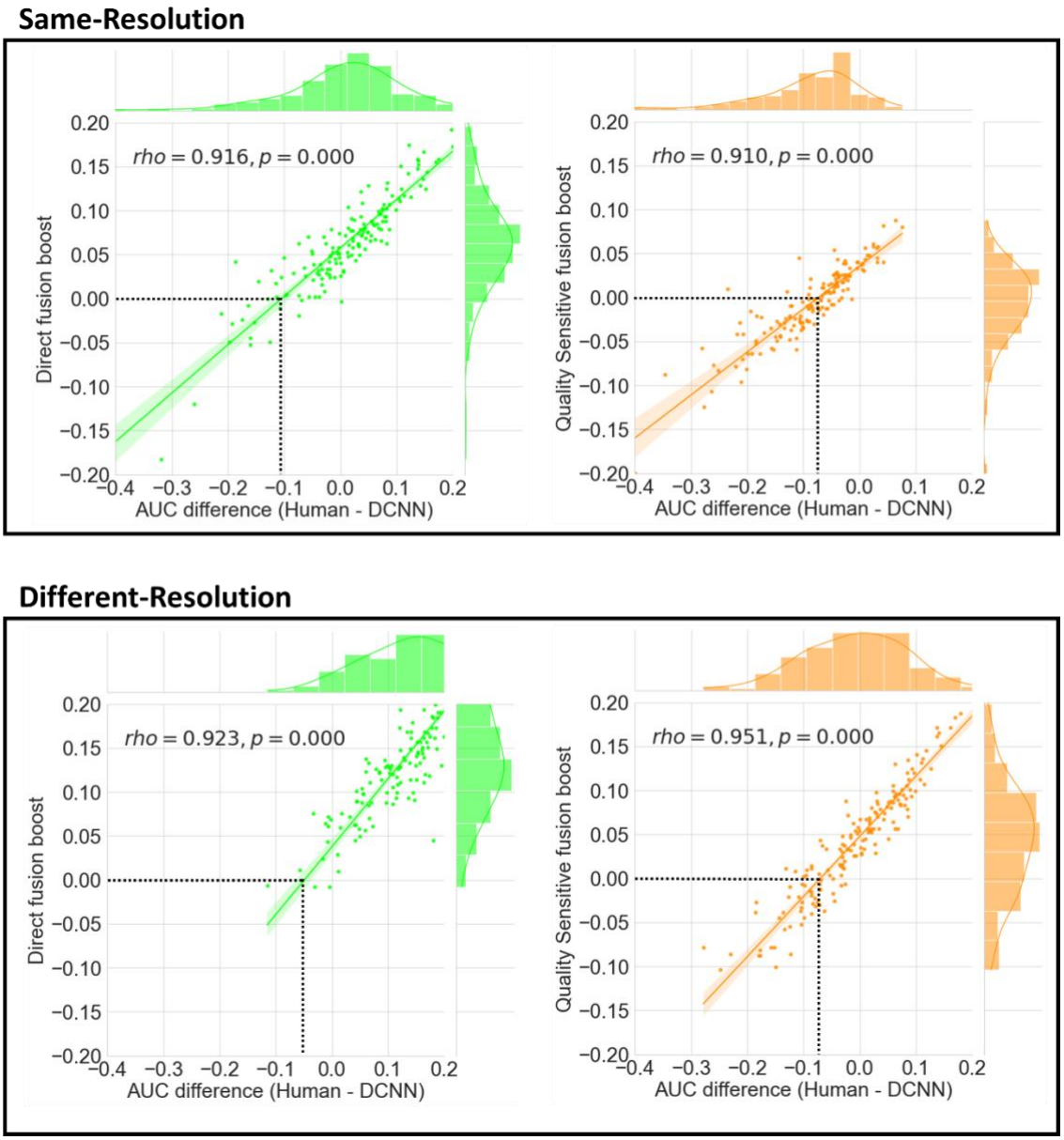


Figure 3.2. Correlation between the difference between human and average DCNN performance (AUC) with their fusion boost in performance using the two fusion metrics. At the top panel, we show the results for the Same-Resolution. At the bottom panel, we show the results for the Different-Resolution condition. Note that we calculate the boost in performance by subtracting the resulting DCNN AUC from the Fusion AUC. Refer to the main text for details regarding the dashed lines.

These significant positive correlations reiterate the previous finding that high-performing humans boost DCNN decisions. However, they also provide the opportunity to specify the level of human accuracy that is required to provide benefit to a DCNN via human-AI teaming. The dashed black lines in Figure 3.2 signify the lowest human accuracy required to boost DCNN performance, i.e. bisecting lines in the vertical axis exactly where

the fusion improvements were zero (AUC_{Boost}= 0). Using the trend line which models our fusion analyses, the projection of such black lines on the horizontal axis shows that humans are valuable for the fusion when their performance is no worse than ~8% lower than the DCNN's absolute performance on average. This result provides a numerical starting value for principled decisions on which humans should be allocated to which DCNNs in human-AI teams. However, DCNN performance varies widely from one algorithm to the next, and so next, we verified the stability of this result for all DCNNs.

Figure 3.3 shows the minimum human-DCNN accuracy discrepancy necessary for fusion to be beneficial to DCNN accuracy, separately for each individual DCNN. For the same-resolution condition, we found that, on average, participants should be no more than below ~-10% of the DCNNs' accuracy to provide a fusion improvement. This value was remarkably consistent across all the DCNNs tested. For the Different-Resolution condition, the average maximum discrepancy was within ~6% on the DCNN, but this was far more variable depending on the individual DCNN that we tested. This variance appeared to be caused by three particular DCNNs (DCNN 7, 8, and 9), while the other 6 DCNNs all showed the same 10% result as in the Same-resolution condition. These 3 DCNNs were also the worst performing, perhaps suggesting that floor effects in DCNN accuracy were responsible for the discrepant results.

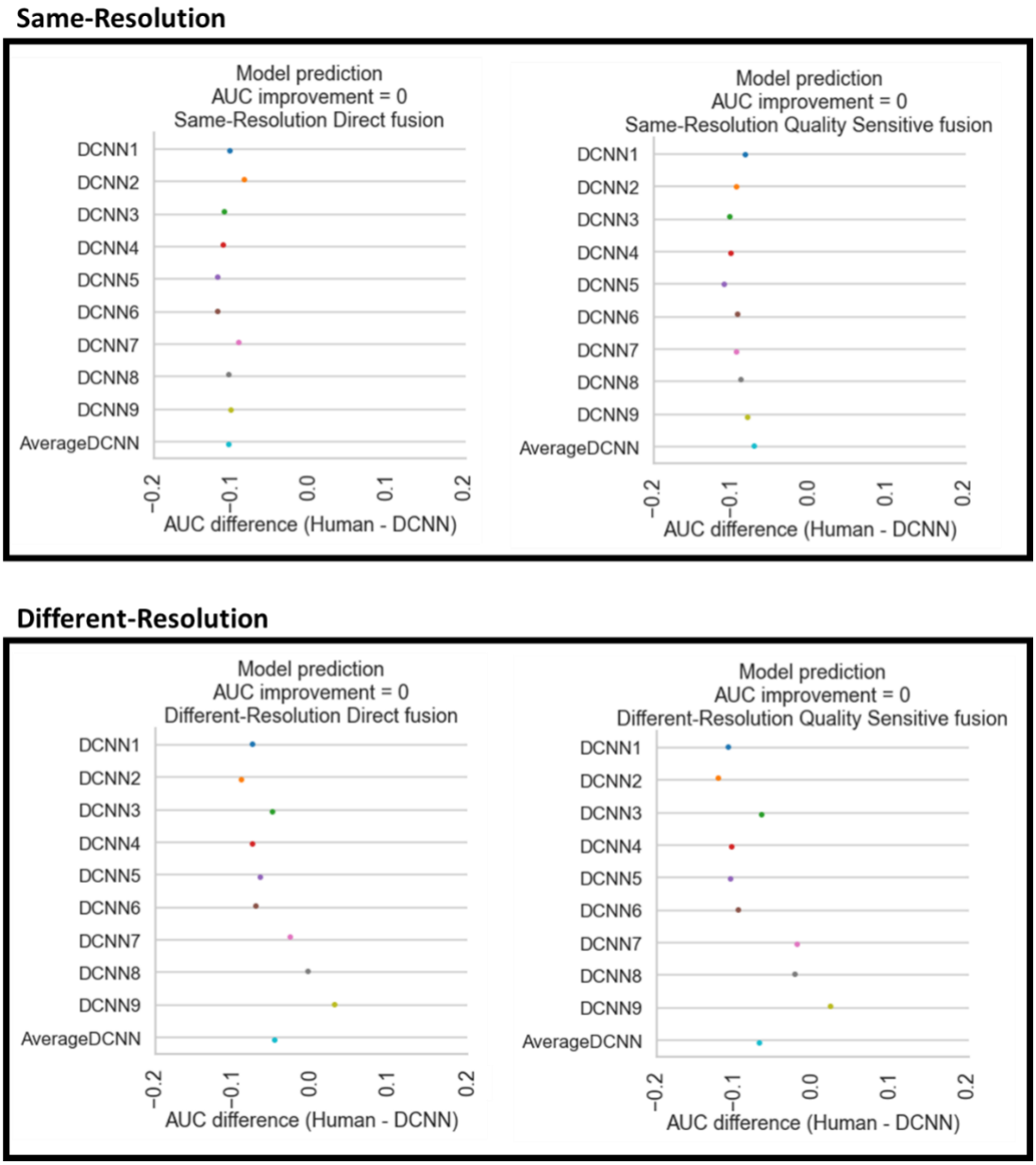


Figure 3.3. The predicted accuracy that humans need to show to improve DCNNs by fusing their decisions when processing face identification. In the top panel, we show the predicted accuracy for the Same-Resolution condition. In the bottom panel, we show the results for the Different-Resolution. For each panel, we show the results for the Direct fusion on the left and the results for the Quality Sensitive fusion method on the right.

Disagreements between humans and DCNN improve fusion performance

Fusion effects arise because uncorrelated errors are cancelled out through response averaging. Therefore we aimed to understand whether performance improvements found in the fusion are related to the level of agreement between humans and DCNNs, as this could be exploited when forming AI-human teams in future. As a reminder, human participants

and DCNNs responded to the same stimuli, and we transformed their responses to a scale from 0 to 1 (see Methods). This scale allowed us to calculate the 1-to-1 agreement (Spearman's correlation) between humans' and DCNNs' responses when processing the stimuli. We predict that further improvements in the fusion effects are related to lower agreement between humans and DCNNs due to more diversity in their responding which would decrease the correlation of errors (see Kittler, Hatef, Duin & Matas, 1998; White, Burton, Kemp & Jenkins, 2013; Hu et al., 2017; Jeckeln et al., 2018).

To examine whether AI-human disagreement was responsible for an additive boost in accuracy in AI-human teams, we first isolated the unexplained variance in the linear models shown in Figure 3.2. As shown in Figure 3.4 (left panels), the AUC difference between individual humans and DCNNs explained the vast majority of variance in the extent to which AI-human teaming boosted accuracy relative to the DCNN alone. However, as shown on the plots on the right, residual variance in these linear models was meaningfully predicted by the degree of agreement between DCNN and human responding.

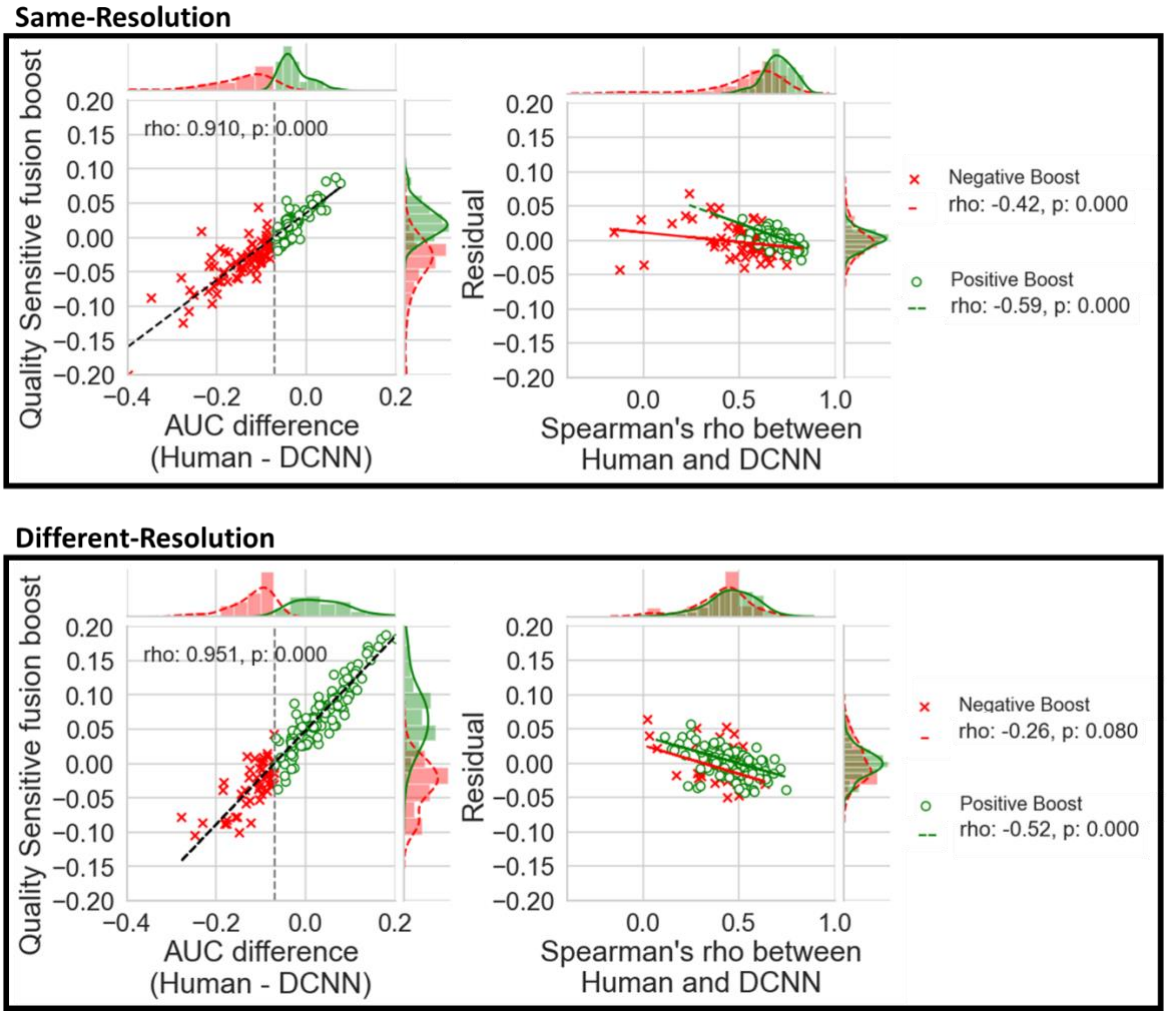


Figure 3.4. Combining humans with DCNNs' has a larger effect when DCNNs and humans disagree on face similarity judgments. On the left of each panel, graphs show the correlation between human and average DCNN performance (AUC) with the boost in performance in the Quality Sensitive fusion metrics. On the right, we show the correlation between the residuals of the linear models on the left with the agreement (Spearman's rho) between humans and DCNNs. For both Same-resolution (top panels) and Different-resolution (bottom panels) image pairs, the boost to DCNN accuracy caused by AI-human response fusion was negatively correlated with the level of agreement.

To carry out the analysis shown in Figure 3.4, we replicated the Quality Sensitive fusion correlation analysis from Figure 3.2 ($AUC_{Difference}$ versus AUC_{Boost}), as shown on the left for each panel in Figure 3.4. We used linear trending regression as a model to predict the minimum accuracy where humans start to improve DCNNs due to fusion. This regression allowed us to determine the exact position where the model predicts the humans that would improve the fusion, dividing the data into expected ***positive*** (green circles) and ***negative*** (red crosses) changes in performance. In addition, because some humans produced fusion effects above - or below – the boost predicted by the model based on

accuracy difference alone, we calculated the residuals (i.e. the vertical distance) between each data coordinate against the model.

The graphs on the right side of Figure 3.4 show correlations between the 1-to-1 agreement of humans and the average DCNN against the found residuals in the Quality Sensitive fusion. We found that the overall correlation between the calculated residual values and agreement (Spearman's rho) between human-DCNN decisions was negative and significant for both the Same-Resolution [$\rho(160) = -0.34$, $p < 0.001$] and Different Resolution [$\rho(166) = -0.40$, $p < 0.001$] conditions. We also separated this analysis into humans who showed positive and negative boosts predicted by the linear model. This analysis showed that humans that improve DCNNs (i.e. green circles in Figure 3.4) showed larger fusion effects for the Same-Resolution [$\rho(160) = -0.59$, $p < 0.001$] and Different-Resolution [$\rho(166) = -0.52$, $p < 0.001$] conditions in comparison to humans that would impair DCNNs (i.e. red crosses in Figure 3.4) in the Same-Resolution [$\rho(160) = -0.42$, $p < 0.001$] and Different-Resolution [$\rho(166) = -0.26$, $p = 0.080$]. Ultimately, these results show that humans who possess higher face processing abilities and rank their decisions differently than DCNNs (i.e. a lower correlation between their 1-to-1 decisions) will improve DCNNs' decisions above the model. This provides a principled basis to make decisions about which humans are suited to be paired with a specific algorithm.

Discussion

In Chapter 3, we re-analysed data from university students and various DCNNs performing challenging face-matching tasks reported in Chapter 2. This computational study aimed to examine the potential for fusing humans and DCNNs to improve the quality of identity verification decisions. We averaged human responses with DCNNs' in face-matching tasks to examine how fused responses improved identity verification decisions compared to the ones made by humans or DCNNs alone. Fusion showed substantial performance improvements for some individuals when paired with some algorithms. However, the extent of the boost provided by humans was highly contingent on the relative accuracy of the individual DCNNs and humans in question.

Understanding that humans can increase the accuracy of DCNNs' decisions is significant because it shows that humans remain a valuable tool for increasing identity verification decisions – algorithms alone are not a complete solution. As mentioned

previously, it is increasingly common for security settings to use automatic recognition systems to perform identity verifications in images taken 'in the wild'. However, as shown in the previous chapter of this thesis (Chapter 2), humans and DCNNs show severe -but different- impairments when processing identity verification in such images.

Perhaps unsurprisingly, we found a positive linear relationship between human performance and enhancements in DCNNs' performance. This result illustrates that humans with higher accuracy in face-matching tests can substantially improve DCNNs' decisions. More interestingly, we were able to identify the level of human accuracy that is necessary to provide a boost to DCNN via response fusion. Using linear regression, we found that humans start to improve DCNNs' decisions if their accuracy is - at a maximum – around 10% lower than the DCNNs' overall accuracy when performing the same stimuli. Curiously, we found that this maximum discrepancy value was very similar across all different DCNNs and stimuli. This robustness across DCNNs allows us to develop a simple model to predict the improvement caused by human-DCNN fusion. Given the linear trend, we found that humans with similar accuracy to the DCNNs ($AUC_{difference}$=0) produced fused identity verification decisions 5% higher than when alone.

Our findings might explain why the work of Phillips and colleagues (2018) achieved ceiling performance when combining forensic examiners - or super-recognisers - with a state-of-the-art DCNN. For example, the DCNN they used showed ~95% accuracy on their test, and the forensic examiners also showed an average of ~93%. Using our results and our practical model, we would predict that fusing a human with similar performance to this DCNN would improve the accuracy of the DCNN by around ~5%, which would make the DCNN achieve ceiling performance.

By showing that human-DCNN fusion effects are proportional to the difference between the facial identification ability of humans and the accuracy of DCNNs, this provides a new benchmark for deciding which humans should be permitted to be part of human-AI teaming in facial recognition. Rather than using arbitrary cut-offs for selecting high-performing humans in face identification tasks (e.g. > +2SD of the mean, see Ramon, Bobak & White, 2019), these results suggest a new principled way of setting these selection cutoffs. Namely, if an individual human performer is able to achieve within 10% of the algorithm accuracy on a face-matching task that is representative of the task performed in

applied settings – or, ideally, above - then they are qualified to contribute face-matching decisions that are combined in the AI-human team.

Interestingly, we also found that the extent to which human responding diverged from DCNNs is also relevant to the improvement that their decisions can bring to a human-AI team. In the context of our study on identity processing, the Likert-type scale allowed participants to express their responses along a spectrum, ranging from strong agreement to strong disagreement, accommodating the nuances and individual differences in how they perceive and process identity-related information. Moreover, use of Spearman's rho allowed us to assess the relationship between human similarity judgments and Euclidean distance computed by facial recognition algorithms (as in previous research: e.g., O'Toole, An, Dunlop, Natu, & Phillips, 2012) without making assumptions about the scaling of this relationship beyond the idea of monotonicity (i.e., that a decrease in Euclidean distance will correspond to an increase in rated similarity, which seems a reasonable assumption). We found that humans who showed less agreement with DCNNs (lower Spearman's correlations) contributed more to the fusion effect. This result is interesting because it illustrates that if fusions are to be applied in real applications, forensic examiners and super-recognisers should work alongside specific DCNNs. That is, a recent body of work from Towler and colleagues (2021) suggests that forensic examiners and super-recognisers use the response scale differently and yet achieve similar accuracy levels (Towler et al., 2021). And so, when applied in real-world settings, the DCNN should - ideally - most disagree with their human pair on ranking identity verification decisions. And so, forensic organisations should count on a body of different DCNNs (i.e. with different architectures, training databases, etc.) and a body of human operators (i.e. super-recognisers, forensic examiners, etc.) to provide rich and diversified solutions. This way, such organisations can determine human operators to work alongside DCNNs specifically 'tuned' to improve their decisions even more. Or even, perhaps, train individuals to rank similarities differently than DCNNs to most benefit fusion effects in face identity decisions. To this end, it will be important to establish in more detail how the Euclidean Distance used by DCNNs relates to the similarity data from the Likert-scale used by humans (beyond the mere assumption of monotonicity). As aforementioned, the appropriate use and interpretation of Likert-type scales—in general—in humans is still a subject of ongoing investigation in the research community.

Establishing the nature of this relationship in the context of the current research would further strengthen the validity of our study.

All the fusion results reported here showed no significant differences between the two metrics for fusion or the filters used to manipulate the quality of the images. This finding is interesting because, despite the considerable difference between the two approaches to calculate DCNNs' similarity scores (i.e. Direct and Quality-Sensitive fusion), the two approaches both produced robust fusion effects. That is, the human-DCNN fusion model was strikingly similar unregarding the metric to calculate DCNNs' similarity scores. Still, as predicted, the Quality Sensitive fusion showed significant performance improvement compared to the Direct fusion, illustrating that clustering images by distinct quality groups somewhat relate to the number of correct decisions made by DCNNs. The observed shift in performance, while maintaining consistent fusion patterns, presents a compelling opportunity to showcase how varied data processing approaches can lead to substantially different outcomes. Simultaneously, this outcome reinforces the robustness of our findings, as both fusion methods continue to yield significant benefits from incorporating human judgment. This is a valuable result because the clustering mechanism of the Quality Sensitive approach prevented DCNNs from being biased in accurately rating specific groups of image qualities. And so, the such approach presents an opportunity for researchers to overcome potential facial verification errors in groups that would offer bias in DCNNs (e.g. race (see Cavazos, Phillips, Castillo, & O'Toole, 2020)) and verify future improvements by performing human-DCNN fusion. Ultimately, our findings demonstrate that robustly trained DCNNs benefit from improved accuracy when integrated with human judgments - as long as humans perform at a similar or higher level than the DCNN. This outcome remains consistent regardless of which of the two metrics is used for human-DCNN fusion. Future studies should try replicate our findings with different metrics and, importantly, test it within and between different groups of participants. This approach would help us better capture the fusion effect in DCNNs and better understand how human decisions contribute to improving its performance. Moving forward, it is crucial to explore the impact of human judgments fused with poorly trained DCNNs and how various datasets sizes can influence human + AI combined performance.

Conclusion

In this work, we demonstrate that combining diverse sources of decision-making can improve face-matching decisions. For that, we investigated multiple setups to understand how fusing independent decisions made by humans and DCNNs can lead to improvements in identity verification. This investigation shows that improvements caused by fusing humans and DCNNs processing the same stimuli have a robust relationship with the accuracy difference between them. We conclude that combining the decisions of humans with similar performance to DCNNs' improves the accuracy of face-matching decisions. In addition, our results show that humans need to perform no worse than 10% lower than the DCNNs' performance to start improving fused identity verification decisions. Also, we show a further accuracy increase proportionally related to differences in the strategies between humans and DCNNs performing the stimuli. Ultimately, Chapter 3 demonstrates a practical use of the differences between humans and DCNNs to improve face-matching decisions.

Chapter 4 - Face information use in humans, super recognisers, and DCNNs

Introduction

Previous chapters compared human and DCNN performance on face-matching tasks. We also examined qualitative differences between DCNNs and humans and how these differences might be exploited to improve performance in human-AI hybrid systems. In Chapter 3, we explored item-level correlations between DCNN and human similarity scores to measure the similarity of face identity representations and the benefit of aggregating human and DCNN face similarity ratings via response 'fusion'. In this fusion analysis, we found that fusing individual humans and DCNNs that were most uncorrelated in their face similarity ratings led to stronger fusion benefits. The implication is that we can exploit individual differences in humans and DCNNs to improve the conjoint performance of human-AI hybrid systems.

However, the perceptual mechanisms underlying these individual differences in face-matching ability still need to be better understood. Therefore, this chapter focuses on better understanding perceptual processing underpinning human and DCNN performance. Our specific aim was to develop novel approaches to eye-tracking analysis that allowed us to examine how human individuals and DCNNs use the visual information in faces to inform face identity decisions.

Individual differences and face information use

Most existing work examining underlying processing mechanisms responsible for individual differences in face perception has focused on the concept of 'holistic' face processing. This is inspired by previous group-level work, which shows that people tend to process faces as a whole (i.e. non-decomposable 'gestalts') rather than by sets of individual facial features (e.g. Farah, Wilson, Drain & Tanaka, 1998; Richler & Gauthier, 2014). The

literature suggests three key image manipulations that appear to demonstrate faces are perceived holistically: The Composite face effect (Young, Hellawell, & Hay, 1987); The inversion effect (Yin, 1969); and the Part-Whole effect (Tanaka & Farah, 1993).

In the Composite Face Task (CFE), researchers align the top half of one face with the bottom half of a different face, creating a new 'chimeric' face identity based on two distinct individuals. This manipulation disrupts participants' ability to recognise the identities in the two face halves. When the face halves are misaligned, participants are then able to recognise the source identities accurately (see Young, Hellawell, & Hay, 1987). Thus, this disruption in identity recognition caused by the chimeric faces argues that identity information depicted in faces is perceived holistically and not as a sum of decomposable facial features.

The Face Inversion Effect (FIE) is the demonstration that face recognition is more impaired by turning the stimulus upside down relative to the recognition of other classes of objects (Yin, 1969). The argument is that inverting faces forces a more piecemeal featural approach due to the disruption in the canonical upright orientation of a face. And so, because of this decrement in performance, the argument is that inverting faces disrupts holistic processing. Finally, the part-whole effect (Tanaka & Farah, 1993) shows that humans are significantly better able to recognise the identity source of single features (e.g. the eyes) when presented in the context of a full face compared to when presented alone (Tanaka & Farah, 1993). This performance enhancement illustrates that humans encode facial features in the context of a holistic representation of a full face.

These findings might suggest that face identity processing is facilitated by a single 'holistic' template used to represent someone's identity. Some studies explored this hypothesis by studying the association between measures of holistic face processing and individual differences in identity processing ability. For example, some studies investigated the relationship between the Composite face effect with measures of face recognition ability (e.g. the classic CFMT (Duchaine & Nakayama, 2006)). These studies mainly found weak correlations between these two measures (e.g. Richler, Cheung, & Gauthier, 2011; Wang, Li, Fang, Tian, & Liu, 2011; DeGutis, Mercado, Wilmer, & Rosenblatt, 2013), and sometimes no correlation with CFMT (Verhallen et al., 2017; Konar, Bennet, & Sekuler, 2009) or other face recognition tests (Richler, Cheung, & Gauthier, 2011; Rezlescu, Susilo, Wilmer, & Caramazza, 2017). In addition, the critical work of Rezlescu and colleagues (2017)

observed the links between face recognition ability with the three main holistic processing measures (see above). Their work shows that only the inversion effect could predict face processing ability. However, most importantly, they report that the three main measures of holistic processing do not correlate with themselves. This vital result shows that, despite measuring what seems to be similar processing mechanisms, these measures alone do not necessarily fully represent holistic processing and do not measure a stable, trait-like individual difference in face processing (see Sunday, Richler, & Gauthier, 2017).

Contrary to the hypothesis that more developed holistic processing mechanisms drive individual differences in identity processing, some recent work shows an association of individual differences with the processing of local part-based information. The ability of participants to perform face identity processing tasks from whole images is highly correlated with their ability to recognise individual and isolated facial features (Sunday, Richler, & Gauthier, 2017). Further, individuals with prosopagnosia show equivalent impairment when processing full or limited facial features suggesting that their impairment is not based on a holistic processing deficit (Tsantani & Cook, 2020).

At the other end of the ability spectrum, super-recognisers require less part-based sampling of face information to maintain accurate recognition compared to typical viewers (Royer, Blais, Gosselin, Duncan & Fiset, 2015). This suggests that their ability enables them to recognize faces from isolated face parts better than typical viewers. Furthermore, image manipulations made to change the global shape of faces (e.g. caricatured faces) do not affect super-recognisers as much compared to other individuals (Kaufmann, Schulz, & Schweinberger, 2013) and manipulations in the spatial layout of facial features affect recognition performance in super recognisers less than typical viewers (Itz, Schweinberger, & Kaufmann, 2018). These pieces of evidence illustrate that holistic processing of face shape and feature configuration processing is less valuable in predicting individual ability than part-based processes (see also Itz, Golle, Luttmann, Schweinberger, & Kaufmann, 2017).

Eye-tracking studies of information use

The research reviewed above underlines the currently limited value of holistic processing measures in explaining individual differences in face identity processing ability. Feature-based processes may therefore appear to be an important avenue for research

aiming to operationalize the processes responsible for differences in face recognition ability. These manipulations could be, for example, by using the classic bubbles technique (Gosselin & Schyns, 2001), where the researcher applies transparent spots to reveal only specific facial regions. Using such image manipulation, researchers found that the average population ranks the eye region as the most diagnostic region for successful identity recognition (Schyns, Bonnar & Gosselin, 2002), and avoidance of this region might relate to poorer face processing abilities (Caldara et al., 2005).

One limitation of the bubbles approach is that it does not enable participants to freely inspect face images as they would usually, and so removes the 'active' component of vision. In this Chapter 4, I use eye-tracking technology to better understand the contribution of this active process of information gathering to individual differences in face-processing ability. Eye tracking devices date from more than two centuries ago (e.g. Huey, 1898), requiring attaching objects to the eyes of the participants while having their eyeballs anesthetised with a - for example- 3% solution of cocaine (see Delabarre, 1898). Today, this technology has evolved to a noninvasive video-based recording, where eye movements can effortlessly be recorded. Recently, this technology has expanded beyond monitoring eye movements when viewing computer monitors to wearable devices that track eye movements in natural settings (e.g. Kassner, Patera, & Bulling, 2014).

Early eye-tracking studies examined participants' attention to scenes that included people and faces (Buswell, 1935; Yarbus, 1965). These demonstrated that people attend to informative regions of scenes, with faces consistently attracting the majority of participants' attention. More recent studies focusing on how individuals distribute their eye fixations on faces show that typical viewers explore the internal features of faces, particularly the mouth, nose, and eyes. This 'T-shaped' fixation pattern is characteristic of viewing across a number of studies (e.g. Henderson, Williams & Falk, 2005; Hsiao & Cottrell, 2008; Iskra & Tomc, 2016; Thomaz, Amaral, Giraldi, Gillies, & Rueckert, 2017; Varela, Ribeiro, Orona, & Thomaz, 2018).

Other work has shown that the standard average gaze pattern found in face identity processing studies varies across individual participants. This variation also appears to vary as a function of face identity processing ability. For example, people with developmental prosopagnosia attend more to external areas of the face, such as the hairline (Avidan & Behrmann, 2021; Stephan & Caine, 2009, but see Lê, Raufaste, Roussel, Puel & Démonet,

2003 for different scan paths between typical viewers and prosopagnosics). Bobak and colleagues (2017) showed images of scenes - containing people – to groups of prosopagnosics, typical viewers, and super-recognisers. Their result shows that prosopagnosics explored faces significantly less than typical viewers, and super recognisers explored the inner features of faces significantly more than typical viewers. Interestingly, time spent on the nose region of faces in scenes moderately correlated with face recognition performance (measured by CFMT (Duchaine & Nakayama, 2006)), showing that super recognisers show a preference to extract information from the centre of faces (see also Bennetts, Mole & Bate, 2017). Varela and colleagues (2018) show that during a face-matching task (i.e. GFMT (Burton et al., 2010)), participants, on average, fixated more on the eye region. Still, those who performed more accurately tended to fixate more on the nose region (Varela et al., 2018). Together, these pieces of evidence paint a mixed picture of the relationship between individual face identity processing ability and eye-movement patterns.

However, the association between greater fixations on the nose region and higher accuracy does appear consistent with other work. For example, Peterson and Eckstein (2012) demonstrate that the face centre is the ideal position for information extraction using a computational approach. They show that the face centre allows maximised amount of facial information extracted compared to other individual facial regions. Focusing on the face centre may therefore be related to holistic processing, allowing a single fixation to extract sufficient information for identification (see also Hsiao & Cottrell, 2008). However, it remains unclear whether the differences between the gaze patterns of super-recognisers and average participants represent more significant holistic processing in super-recognisers. In a study using a gaze-contingent eye-tracking paradigm where participants viewed faces through 'spotlight' apertures, Miellet and colleagues (2013) show that the preference for central fixations is observed even when participants could only see a small part of the face at one time. This suggests that central fixation does not necessarily denote holistic processing.

Our approach

Eye-tracking has been a valuable tool for investigating information sampling in face processing. The high data collection, speed, and precision of such machines allow

researchers to examine how humans differ when performing, for example, a face-matching or recognition task. However, the work described above provides a somewhat mixed picture that does not reveal the fundamental differences in information sampling that underpin differences in face identification ability.

Here, we develop novel approaches to eye-tracking analysis to deepen our understanding of the link between face information sampling and differences in face processing ability. Prior studies typically analyse the proportion of eye fixations inside delimited Regions-Of-Interest (ROI) (e.g. nose, mouth, etc.). As shown in Figure 4.1, using heatmaps enhances the resolution of the investigation by using all pixels of the image instead of a few delimited regions. But heatmap analysis poses technical problems in making statistical comparisons (see Lao, Miellet, Pernet, Sokhn, & Caldara, 2017), and these same problems somewhat limit the use of heatmaps in exploratory studies of individual differences. In my master's work, I applied a data-driven approach to characterize how humans sample information from faces (Varela et al., 2018). Instead of counting the number of eye fixations by delimited regions and their respective dwell time, I used methods based on statistical learning to analyse heatmaps of fixations distributed continuously across the face.

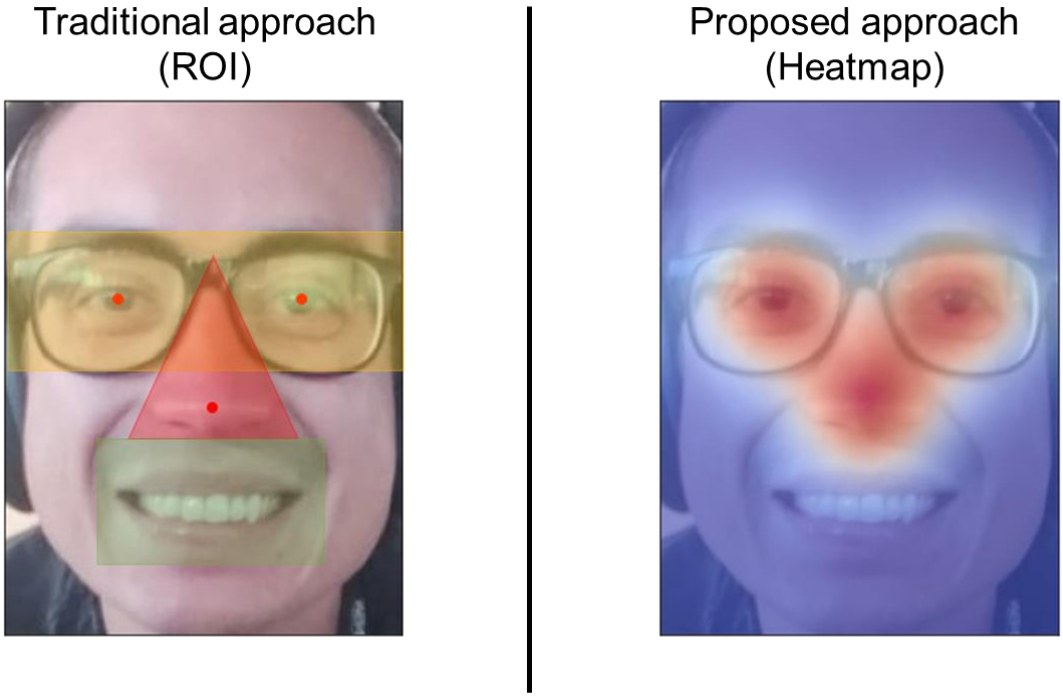


Figure 4.1. Examples of the different approaches to investigating the eye strategy of humans by eye fixation coordinates. In this figure, two fixations are made to the eye region and one to the nose. We show the ROI on the left panel and the heatmap representation on the right. In this chapter, I develop a new approach to analysing heatmap data.

Varela and colleagues (2018) key novelty was to use of a linear predictor (PCA+MLDA: Thomaz, Kitani & Gillies, 2006) to compress an array of salient map images into a single dimension. By using images large in resolution (e.g. multiple thousand pixels), it is often difficult for supervised learning algorithms to fit the given information – as, for example, the number of human datapoints in most scientific experiments using human data and images is lower than the resolution of the images. In their experiment, Varela et al's (2018) approach compresses few-N salient map images of relatively *large resolution (e.g. more than 10000 pixels)* to a reduced numerical value for subsequently classifying face recognition ability using the MLDA. However, despite showing incredible classification accuracy in discriminating high versus low performers using this technique, their study was preliminary and did not visualise the dimension of variance in salient maps that explained inter-individual differences. Therefore, we aim to extend their approach and visualise what this dimension shows to provide clues as to the strategic differences in information sampling behaviour that give rise to individual differences in face identity processing ability.

Chapter Objectives

This chapter reports three experiments designed to investigate whether individual differences in face identity processing ability can be explained by differences in the information that is sampled from a face. The first experiment (Experiment 2) will analyse participants' fixation patterns while performing a face-matching task using PCA and MLDA analysis. This experiment will help us understand the facial features associated with higher accuracy when requiring no memory component. Experiments 3A and 3B then examine participants' fixation patterns using a similar approach but in a face recognition task, which relies on participants memorising and later recognising faces. Here, 'super-recognisers' – people with high levels of ability in face identification – are compared to typical viewers when performing a gaze-contingent 'spotlight' procedure where faces are viewed through

apertures centred on the fixation of the participant. This approach enables more precise control over the face information that is being viewed by participants on a given fixation, enabling the information sampled by super-recognisers and typical viewers to be compared.

Finally, Experiment 4 will use a novel approach to quantify the computational value of face information sampled by human participants in Experiment 3. By providing DCNNs with the samples of information extracted by participants in Experiment 3, we will ask whether human-led information sampling leads to greater face identification in DCNNs compared with random samples of information. Some recent studies suggest that humans and DCNNs make use of similar facial features to humans making decisions regarding someone's identity (Abudarham and Yovel, 2016), leading to the basic prediction that information sampled by humans would contain more identity information. We also ask whether the computational value of information Is sensitive to individual differences in viewers' face identity processing ability by comparing DCNN accuracy with face information sampled by super-recognisers and typical viewers'. If the information sampled by super-recognisers contains more computationally useful face identity information, this would point to the importance of information sampling in explaining individual differences in face recognition ability.

Experiment 2

Experiment 2 investigates participants' fixation patterns while performing a self-paced face-matching task. Two face images are presented side-by-side, and participants must make a binary decision regarding the identity of the two faces (i.e. same person or different people) (e.g. Burton et al., 2010). To our knowledge, there are a relatively small set of studies examining participants' fixation patterns while performing face-matching tasks (Havard, 2007; Varela et al., 2018; Özbek & Bindemann, 2011). Together these studies provide a mixed picture of information used in face-matching tasks. The studies of Varela and colleagues (2018) and Havard (2007) show that the eye region is the most observed amongst typical viewers. However, the eyes alone do not provide enough information for accurate matching (Havard, 2007), and better ability relates to more observation of the nose region (Varela et al., 2018).

Importantly, in these studies, participants had unrestricted time to make their decisions regarding the identity depicted on two facial images (see Megreya, Bindemann, Havard, & Burton, 2012 for a study containing more facial images per trial). Özbek and Bindemann (2011) examined the fixation patterns - and accuracy - of participants performing a matching task under random time restrictions (200, 500, 1000, or 2000ms). Their results show that the initial eye fixations landed mostly on the face centre (i.e. nose) in all restriction cases, suggesting either that the central region is optimal for visual efficiency when there is little opportunity to explore the face because it is the centre of mass of the object, or both (Özbek & Bindemann, 2011; but see Bindemann, 2010).

Performance on unfamiliar face-matching tasks is thought to be more reliant on part-based processing when compared with face recognition tasks (Megreya & Burton, 2006). If this is the case, we might predict that attention to features containing high levels of identity information would be fixated on these tasks. For example, some studies have shown greater face identity information contained in the eyes (Bate, Haslam, Tree & Hodgson, 2008; Schyns, Bonnar & Gosselin, 2002; Slessor, Riby & Finnerty, 2013), while others show strong identity cues in the ear region (Towler et al. 2017, 2021). Alternatively, concentrations of fixations on the nose could support high performance if holistic processing is associated with face identity processing skills (Bobak et al., 2017; Bennetts, Mole & Bate, 2017; see Wang et al., 2012).

In Experiment 2, our objectives are two-fold: (i) introduce novel computer vision and statistical learning techniques as an exploratory analysis to visualise - and comprehend - the fixation strategies inferring face processing ability; (ii) to visualise and investigate fixation patterns associated with high performance in face-matching tests. We collect the fixation patterns of a small cohort of university students performing a standard face-matching test using an eye tracker. This data is analysed using methods introduced by Varela and colleagues (2018), but here we will visualise the virtual dimensions found by the PCA and MLDA analysis. In addition, we will also analyse the data using the standard Region-of-Interest approach to examine whether it provides complementary information to our new analysis method.

Methods

Participants

A total of 44 university students participated in this study in exchange for university credits. We collected the data from 34 females and 10 males, whose average age was 19.15 years (SD = 1.82).

Stimuli

Participants performed the Glasgow Face Matching task (GFMT) (Burton et al., 2010) short version. The GFMT is a standard test for measuring face-matching ability. The GFMT is a challenging face-matching task containing 40 randomly presented face pair trials in grayscale. Of these images, 20 are from different people, and 20 are from the same person. Participants processed these face pairs making a binary decision regarding the identity of the two faces of the pair while the eye-tracking equipment measured their eye positions. Participants had no time constrain to process the face images.

Procedure

Participants completed the GFMT on a Laptop screen measuring 38 by 22 cm. We recorded participants' eye gaze data with an SR-Research EyeLink Portable Duo (with a chin rest) tuned to sample data at 1000hz. We sat participants in front of the monitor and eye tracking device at a distance of 55 cm from the monitor so that the angular size of the screen was ~38° of visual angle (each face of the stimuli covered ~12° of visual angle, or 12 cm). This device has an average gaze position error of about 0.25° and a spatial resolution of 0.01°. Before data collection, we conducted a calibration procedure for eye fixation using a nine-point fixation procedure programmed in MATLAB, repeating the process until reaching a satisfactory alignment. We only tracked participants' dominant eye data and presented the stimuli in MATLAB using the EyeLink toolbox (Cornelissen, Peters & Palmer, 2002).

Eye movement classification

We classified eye gaze data into eye fixations and saccades. We classified saccades based on a velocity threshold above 30° of visual angle/sec. We coded adjacent samples below the velocity threshold as eye fixations and adjacent samples above the velocity threshold as saccades between fixations. We only used eye fixation data for this experiment.

We cleaned eye fixations based on three parameters: (i) removed fixations shorter than 50ms; (ii) removed fixations longer than three standard deviations from the mean fixation duration; (iii) removed fixations that landed outside of the face. Importantly, we aimed to remove trials with no valid fixations and remove the participant if we excluded more than 20% of their trials. We excluded only one trial of one participant due to no valid fixations.

Analysis of fixation patterns

We analysed the fixation data using two distinct approaches. First, we analyse fixation patterns as heatmap images using the PCA+MLDA approach introduced by Varela et al. (2018). Second, we analyse fixation patterns using a standard regions of interest (ROI) approach. This enables comparison with our novel PCA+MLDA visualisation approach to assess whether it provides reliable insight into the facial information supporting high accuracy in face identification tasks.

Principal components analysis (PCA) of Heatmaps

We first transformed a set of fixation coordinates into a three-dimensional histogram (heatmap) image using iMap4 (Lao et al., 2017) in MATLAB. This procedure is important because iMap4 smooths the fixation data using Gaussian kernels, which could represent realistic attentional constraints in the surrounding areas adjacent to the fixation coordinate. This procedure transformed an array of fixation coordinates into a three-dimensional image containing the same resolution as the stimuli. Each pixel of this heatmap image received a score proportional to the attention received. We processed the data of all participants and normalised each heatmap individually.

We compressed heatmaps of all participants using Principal Components Analysis (PCA). We used this exploratory procedure to visualise the main sources of variance in fixation heatmaps. The PCA transforms *N* three-dimensional heatmaps into *m* unique ($m \leq N$) Principal Components (PCs, or eigenvectors), where each heatmap receives a projected loading score within each PC. These PCs are new uncorrelated variables constructed from linear combinations of the initial heatmaps, representing the fixation pattern differences in *m* possible directions ranked by the amount of explained information (i.e. variance or eigenvalues). For the purpose of our analysis, the PCA enabled us to visualise the main sources of variance that distinguish the viewing patterns of participants in our study.

To conduct a PCA on heatmaps, we first resized the original heatmap images from *1920x1080* to *250x175* pixels and converted each to a single-dimension vector (i.e. one heatmap = 1x43750 pixels). We reduced the image size for less computational effort. Then, we normalised these vectors and applied the PCA to the transformed *Nxn* matrix*,* where *N* represents the number of heatmaps, and *n*, the concatenated pixel information contained in the image (43750 pixels). We visualised these PCs as their dimension interacting with the average heatmap. To observe if variations in heatmaps captured by PCs were associated with face-processing ability, we correlated the loading scores of PCs with face-processing ability measured by the stimulus.

In the analysis that follows, we performed PCA using both the participant-level average heatmaps and also the trial-level maps. This two-level approach was chosen given the exploratory nature of this research. Participant averages enabled us to focus on inter-individual differences in face information use, whereas trial-level heatmaps enabled us to capture the intra-individual variation, i.e. the variation in individual participants' fixation patterns over different image pairs.

Maximum uncertainty Linear Discriminant Analysis (MLDA) of Heatmaps

The PCA will significantly reduce the quantity of data to analyse in the experiment and enable us to visualize the main sources of variance. However, one of our objectives with this study is to visualise an optimal eigenvector that represents the main source of inter-individual variation in heatmaps that *explains* face processing ability. For that, we applied a Maximum uncertainty Linear Discriminant Analysis (MLDA) (Thomaz, Kitani & Gillies, 2006)

to transform the *Nxm* matrix from the PCA to an *Nx1.* The MLDA essentially finds the optimal eigenvector within the data that best separate groups of individuals. Consistent with earlier work (Varela et al., 2018), we assigned the top 10 participants to the 'Higher Ability' group and the remainder to 'Average Ability'. Visualising the dimension of variance in heatmaps that best discriminate these groups might clarify the information sampling strategies that result in accurate face processing. Figure 4.2 shows a schematic illustration of the computational framework used in the MLDA analysis.

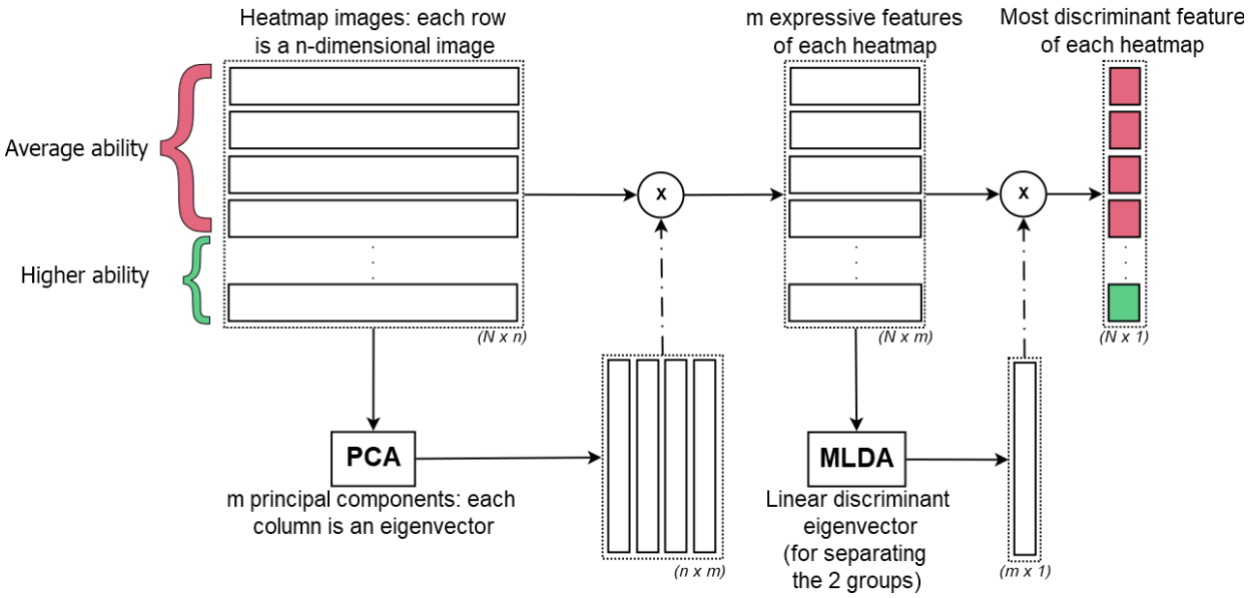


Figure 4.2. Schematic representation of the PCA+MLDA process used to identify face information associated with face processing ability. The PCA transforms the data into compact vectors, and the MLDA finds the optimal eigenvector for dividing groups of participants based on their face processing ability (see Thomaz, Kitani & Gillies, 2006).

Regions-Of-Interest

To support the PCA+MLDA analysis, we also investigated the eye gaze patterns of participants distributed across Regions-Of-Interest (ROI). The ROI analysis is the most common method of eye gaze analysis found in the literature to investigate facial features (e.g. Bate, Haslam, Tree & Hodgson, 2008; Slessor, Riby & Finnerty, 2013; Bobak et al., 2017; Bennetts, Mole & Bate, 2017). This method allows insights from the visualisation generated

by the PCA - or MLDA- to be compared with an ROI analysis to establish this approach for the study of individual differences in face identity processing.

To delimit facial regions, we first computed the average stimulus. The GFMT (Burton et al., 2010) is a face-matching task in which all 40 face pairs are vertically aligned and posted side by side on the computer screen. Therefore, we used the average image across all 40 trials as a model to delimit five facial ROIs for the two faces: Left eye, right eye, between eyes, nose, and Mouth (See Figure 4.3). We collected fixations that landed within these regions for both faces of the pair (e.g. fixations on the left eye are the sum of the fixation on the left eye for both faces). We report the proportion of fixations and dwell time within each ROI.

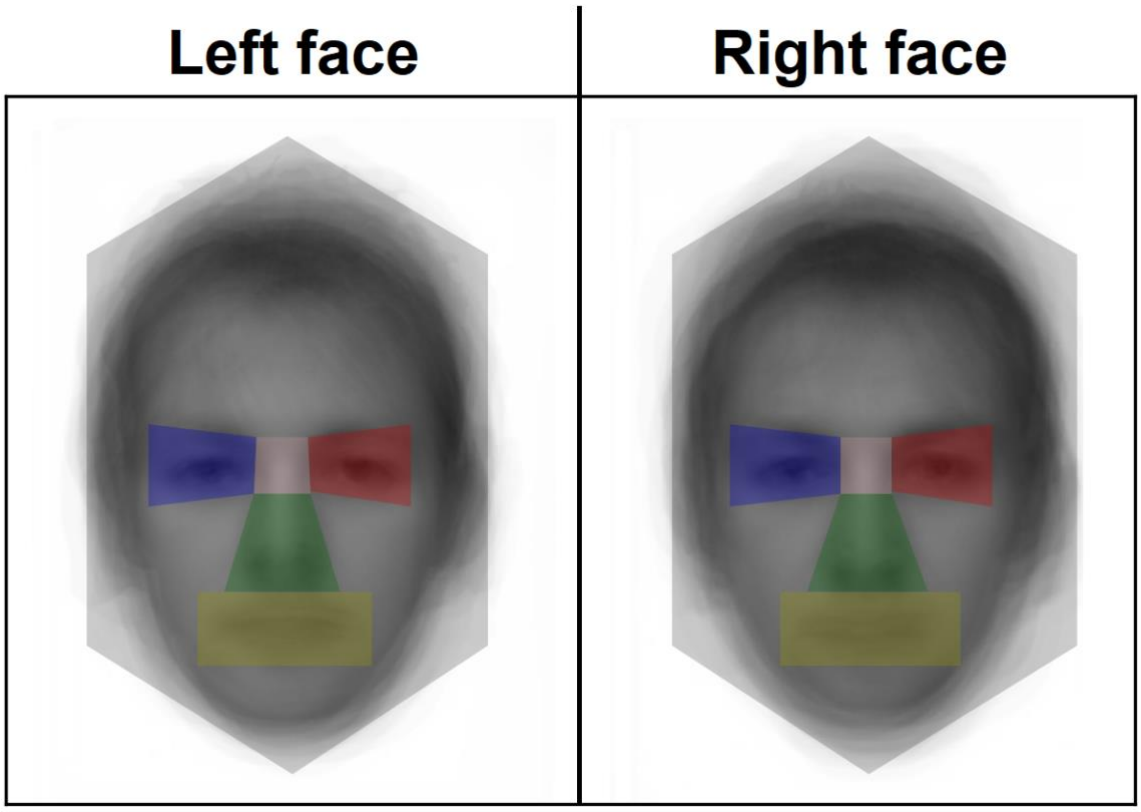


Figure 4.3. Facial regions on top of the average stimulus. Here, we show the model for ROI that we created to investigate patterns of fixations in participants. We divided the two faces of the pair into five possible facial regions: Left eye (blue), right eye (red), between eyes (pink), nose (green), and Mouth (yellow).

Gini coefficient as a measure of gaze dispersal in heatmaps

The PCA+MLDA and the ROI analysis will give us cues regarding fixation patterns and localised facial regions participants attended to process the stimuli. However, these analysis methods will not directly measure the overall dispersion of participants' fixation patterns. The literature mentions that face processing ability significantly correlates with time spent observing the central facial area (e.g. Bobak et al., 2017; Bennetts, Mole & Bate, 2017) and that super recognisers require less information for accurate recognition (Royer et al., 2015). Thus, it is possible that individuals with enhanced face processing would conduct less visual exploration of faces, so we also measured the amount of exploration our participants engaged in when performing the stimuli.

We used the Gini coefficient to measure the dispersal of fixation patterns (Lorenz, 1905). The Gini coefficient represents inequality among values of a given distribution. Therefore, when applied to heatmaps, the Gini coefficient would quantify the amount of dispersion (i.e. exploration) that participants engaged when processing the stimuli. Higher Gini coefficients would represent concentrated fixation patterns, and lower Gini coefficients would represent highly dispersed fixation patterns.

Results

Overall Accuracy

Forty-four participants completed the 40 pairwise face-matching decisions from the GFMT (Burton et al., 2010). The average score, measured by correct hits, was 78.7% (SD= 11.51%) compared to the average accuracy of 81.2% (SD= 9.47%) in the original publication of the GFMT (see Burton et al., 2010). Differences between our results and the original normative accuracy were not significant [t(236)= 1.54, p=0.126]. To divide between groups of performance, we established that the 34 worst performers were part of the Average ability group (M=74.3%, SD=9%), and the top 10 were part of the Higher ability group (M=93.8%, SD=1.77%).

Comparison of heatmap and ROI approaches

We show the proportion of fixations for each ROI and the average fixation pattern as a heatmap in Figure 4.4. On the left panel of Figure 4.4, we show the average proportion of time spent on all ROIs. According to the ROI data, participants spent 61.7% of their total fixations observing the face pair's five delimited inner facial features. Amongst these features, the nose was the most observed, followed by the eyes and mouth.

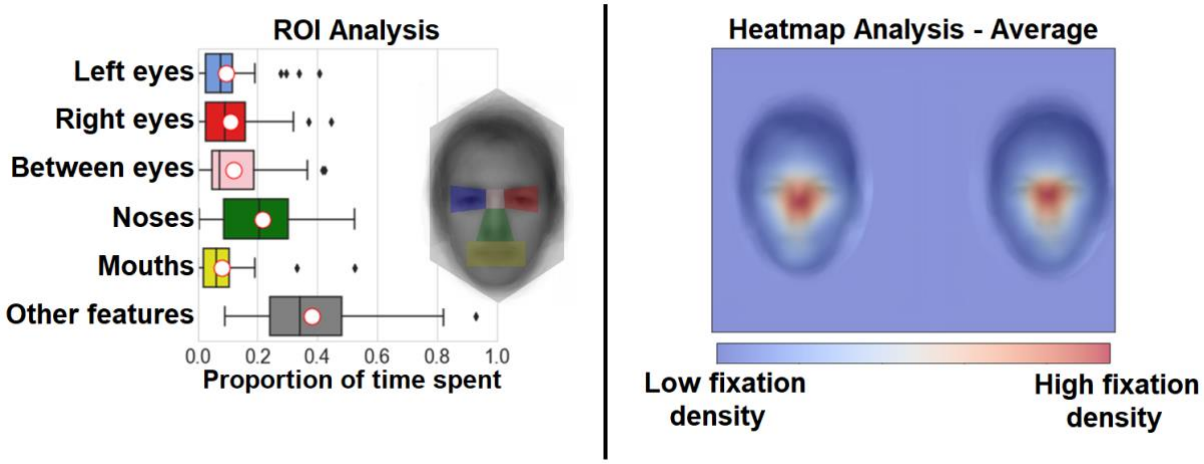


Figure 4.4. Results for the average gaze of participants using the two analysis methods used in this study. We plot the participant-level ROI analysis on the left side of the panel, showing the proportion of time participants spent fixating on delimited ROIs. We show the resulting average heatmap from these fixations on the right side of the panel.

On the right panel of Figure 4.4, we show the density of fixations by plotting the average fixations heatmap. Consistent with the ROI analysis, the average fixation heatmap reveals a higher density of fixations on the central area of the faces, spreading further to the eyes, between eyes, and mouth. The ROI analysis shows that, on average, 38.3% of participants' fixations landed on different face elements other than the five ROIs (e.g. cheeks, hair, chin, etc.). And yet we do not observe clusters of fixations on the surrounding areas of the face in the heatmap analysis. One possible explanation for this effect is that these fixations landed on different external feature regions for particular trials, and because the heatmaps are averages across trials, this resulted in only the common fixations being visualised. Another explanation could be that the delimitations of ROIs were too narrow to collect all fixations landing in these areas, given the margin of uncertainty in the precise fixation locations. Therefore, a portion of 'other feature' fixations could have resulted from

fixations that landed near bounding areas outside ROIs. This underlines the importance of applying multiple analysis techniques to enable converging evidence on which stronger conclusions can be based.

Inter-individual difference analysis of fixation patterns

We performed a PCA on the participant-level heatmap data to explore the individual differences in how people disperse their gaze across face images and its relation to ability. In Figure 4.5, we ordered the first principal components (PCs) from top to bottom by the amount of variance explained by them (total variance = 64.8%). Visual analysis of Figure 4.5 allows us to understand better what each principal component represents, and it appears that some PCs offer semantically explainable differences in fixation patterns across participants. For example, the first PC appears to represent an inter-individual difference whereby some participants attend to the left of the faces and others to the right. PCs 2 and 3 both appear to show more focal attention on the central and nose regions on the left versus more diffuse attention across the upper facial features on the right.

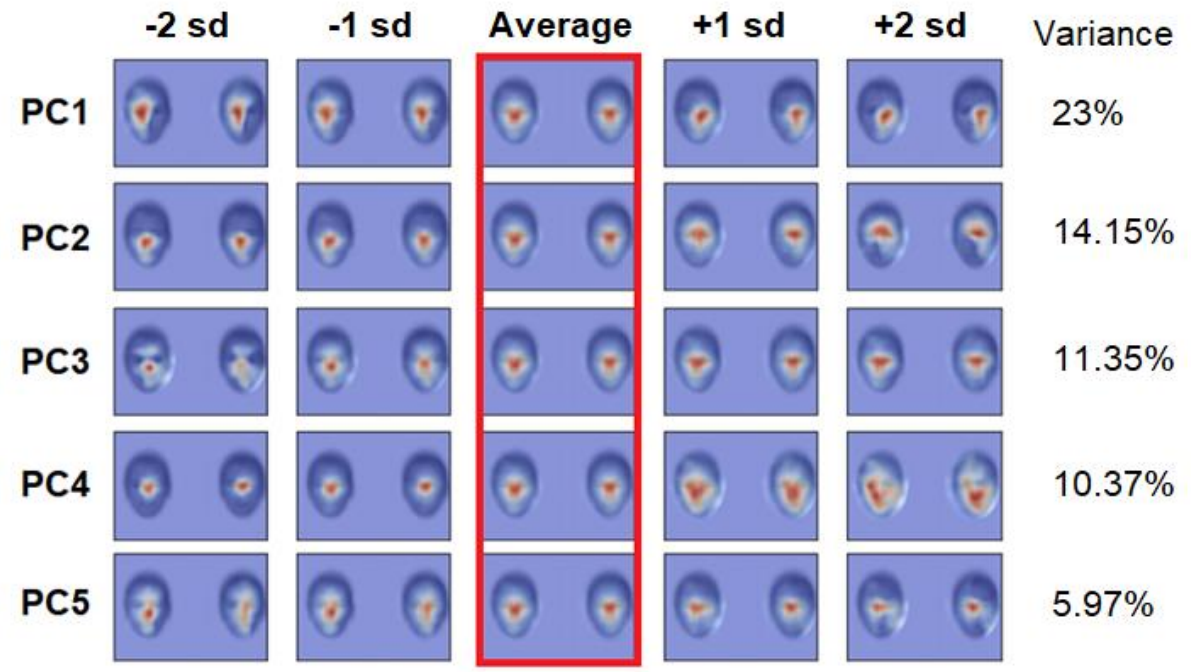


Figure 4.5. Visualisation of the first five principal components in the heatmap analysis of Experiment 2. This figure shows PCs obtained using the participant-level heatmap data and their corresponding explained variances. In the PCA, the average heatmap receives a loading score of zero for each PC, and a zero-mean normal distribution represents participants' loading scores. Therefore, some participants received a negative loading score (i.e. to the left side of the average). And some received a positive loading score (i.e. to the right side of the average). Visual inspection allows some inferences about major sources of individual differences. For example, PC1 shows a facial side preference to which people attended to the face pair. PC2 shows that some people investigated the nose and mouth region while others attended more to the eyes.

Next, we calculated the loading score for each participant on each of the 5 PCs in Figure 4.5. These five loading scores for each participant provide a profile that describes how their fixation patterns related to the overall variance in fixation patterns observed across all participants. Participants receiving a negative loading score for a given PC indicate a fixation pattern more characteristic of patterns to the left side of the average, and positive loading scores are more characteristic of patterns to the right side of the average.

Figure 4.5 visually describes the five first PCs for the participant-level heatmap data. However, one problem that arises from this visual analysis is that it relies solely on subjective interpretation. So, to support the PC investigation in Figure 4.5, we performed a correlational study of PC loading scores with ROI fixations. This comparison is shown in Table 4.1 and helps to corroborate the subjective approach.

		Left eyes	Right eyes	Between eyes	Noses	Mouths	Other features
PC1	Spearman's rho	-0.25	0.283	-0.13	-0.03	0.28	0.018
	p-value	0.11	0.062	0.394	0.874	0.066	0.906
	Confidence Interval	[-0.509099 0.050644]	[-0.015159 0.534941]	[-0.411022 0.173585]	[-0.324000 0.269284]	[-0.018416 0.532611]	[-0.280384 0.313212]
	Spearman's rho						
	p-value	0.714	<.001	0.108	0.001	0.009	0.709
	Confidence Interval	[-0.350640 0.241182]	[0.254943 0.702841]	[-0.054895 0.505935]	[-0.672980 -0.201187]	[-0.615607 -0.105308]	[-0.243071 0.348878]
PC3	Spearman's rho	0.252	0.271	0.299 *	0.067	0.159	-0.122
	p-value	0.099	0.075	0.049	0.667	0.302	0.429
	Confidence Interval	[-0.048515 0.510679]	[-0.028150 0.525598]	[0.002320 0.547302]	[-0.234551 0.356789]	[-0.144717 0.435336]	[-0.404244 0.181458]
PC4	Spearman's rho	-0.32 *	-0.19	-0.29	0.06	0.274	0.111
	p-value	0.036	0.227	0.053	0.698	0.072	0.471
	Confidence Interval	[-0.563364 -0.025541]	[-0.460888 0.113275]	[-0.540362 0.007534]	[-0.241182 0.350640]	[-0.024912 0.527939]	[-0.192220 0.394873]
PC5	Spearman's rho	0.393 **	0.131	0.309 *	-0.05	-0.18	-0.091
	p-value	0.009	0.396	0.042	0.731	0.242	0.557
	Confidence Interval	[0.108810 0.617803]	[-0.172598 0.411867]	[0.013338 0.554974]	[-0.341812 0.250606]	[-0.452694 0.123484]	[-0.377682 0.211602]

Note. * p < .05, ** p < .01, *** p < .001

Table 4.1. Correlational analysis of Principal Component loading scores with fixations in ROI.

For example, PC 2 loading scores had significant negative correlations with nose and mouth fixations and a positive correlation with the right eye fixations. These significant results are somewhat correspondent with the subjective impression that this PC was indexing the degree of central processing. While some of the other patterns do not reach statistical significance, the correspondence nevertheless lends support to our subjective interpretation, with PC1, for example, positively correlating with right eye fixations (p = 0.062) and negatively with left eye fixations (p = 0.11).

Previous studies have found associations between fixation patterns and face processing ability (e.g. Bobak et al., 2017; Bennetts, Mole & Bate, 2017, Varela et al., 2018). To examine this, here, we correlated the five first PC loading scores with face processing ability as measured by overall accuracy scores in the GFMT. We found that loading scores for PC1 [rho(42) = 0.3, p = 0.048, CI = [0.0034 0.5481]] and PC3 [rho(42) = 0.46, p = 0.002, CI = [0.1889 0.6659]] significantly correlated with performance. This result indicates that participants showing enhanced face-matching ability explore the right side of the face pair (i.e. positive loading score in PC1) and tended to distribute fixations more on the eye region of a face (i.e. positive loading score in PC3). We replicated this analysis for ROI fixation patterns and found that GFMT performance was associated with more fixations landing on the right eyes [rho(42)=0.32, p=0.036, CI = [0.0255 0.5634]]. Full statistics from these correlational analyses are shown in Table 4.2.

Correlation Matrix
PCs vs GFMT Score

		GFMT Score	
PC1	Spearman's rho	0.3	*
	p-value	0.048	
	CI	[0.0034 0.5481]	
PC2	Spearman's rho	0.16	
	p-value	0.3	
	CI	[-0.1437 0.4362]	
PC3	Spearman's rho	0.46	**
	p-value	0.002	
	CI	[0.1889 0.6659]	
PC4	Spearman's rho	0.1	
	p-value	0.539	
	CI	[-0.2029 0.3854]	
PC5	Spearman's rho	-0.12	
	p-value	0.437	
	CI	[-0.4025 0.1834]	

Note. * p < .05, ** p < .01, *** p < .001

Correlation Matrix
ROIs vs GFMT Score

		GFMT Score	
Left eyes	rho	-0.1	
	p-value	0.523	
	CI	[-0.3854 0.2029]	
Right eyes	rho	0.317	*
	p-value	0.036	
	CI	[0.0222 0.5611]	
Between eyes	rho	0.054	
	p-value	0.73	
	CI	[-0.2468 0.3453]	
Noses	rho	-0.15	
	p-value	0.326	
	CI	[-0.4278 0.1537]	
Mouths	rho	0.148	
	p-value	0.337	
	CI	[-0.1557 0.4262]	
Other features	rho	0.04	
	p-value	0.799	
	CI	[-0.2600 0.3329]	

*Table 4.2. Tables showing how Principal Components (PCs) and fixations on ROIs correlated with face-matching ability measured by the stimuli (GFMT: Burton et al., 2010). We show Spearman's correlation between the five first PCs against GFMT score on the left table. And show Spearman's correlation between fixations in ROIs against GFMT score on the right tabl*e.

Intra-individual difference analysis of fixation patterns

Although participant-level heatmaps allowed our inter-individual difference analysis to focus on information sampling between individuals, we also observed substantial variation between fixation heatmaps on individual trials for each participant. To further explore this, we examined each participant's intra-individual variation in fixation patterns. We show an example of inter-trial variability in Figure 4.6, with detailed results for this analysis in APPENDIX C.

In Figure 4.6, we plot PC1 loadings for the trial-level heatmap data separately for each participant. The stimuli had 40 trials, so stability in PC scores would convey stability in information sampled by eye fixations across trials. However, an inspection of Figure 4.6 reveals that the information sampled varied substantially for each participant. Quantifying this variation, the average range of participants' PC1 scores (*range = max - min*) was of 2.45 standard deviations, although this variability did not appear to be related to participant ability (Average ability group = 2.39; Higher ability= 2.66; $t(42)=-1.21$, $p=0.232$). Nevertheless, this result is important because it reveals that participants show a high degree of flexibility in sampling facial information and, here, independent of face-matching ability. So while an average heatmap captures the consistency in participants' fixation patterns across trials, it removes an important source of variation in the information that is sampled in any given trial.

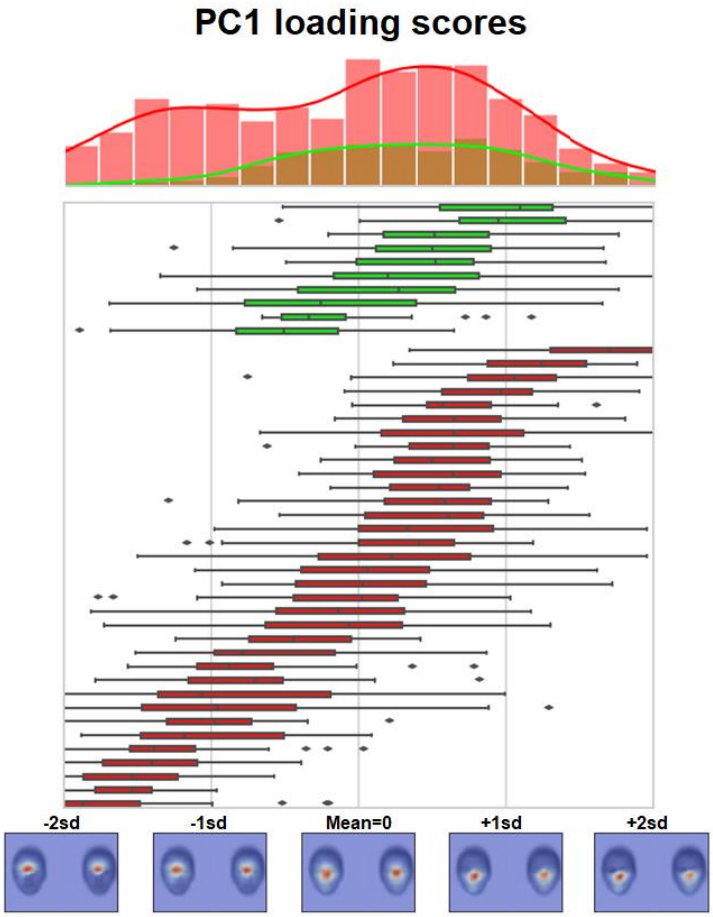


Figure 4.6. Intra-individual variation in loading scores across GFMT trials for Principal Component 1 (left). Although only PC1 is shown here, intra-individual variation is similar in magnitude for all PCs. The boxplots show the interquartile range, and the whiskers show the max and minimum of each distribution. The red bars stand for participants in the Average ability group. The green bars stand for participants in the Higher ability group. Participants' data are ordered from top to bottom based on the mean loading scores of each group but show striking intra-individual variation in all participants.

Individual difference analysis of heatmaps using MLDA

We used a Maximum uncertainty Linear Discriminant Analysis (MLDA: Thomaz, Kitani, & Gillies, 2006) to find the best eigenvector within the PCA space to discriminate between 'high-performers' and other participants. For that, we divided participants into two distinct categories of performance, called Average and Higher ability groups. We arbitrarily divided the group of participants into the bottom 34 performers (Average ability group: M=74.3%, SD=9%), and the top 10 performers (Higher ability group: M=93.8%, SD=1.77%). The MLDA, then, found the eigenvector that best separates the fixation patterns between

these two groups. The aim was to visualise the pattern of fixations that distinguish high performers from other participants.

We visualise the discriminating MLDA eigenvector in the top panel of Figure 4.7 as a heatmap image. Subjectively interpreting this heatmap image reveals that participants of the enhanced performance group (i.e. greener regions in the heatmap) showed fixations predominantly on the right side of the faces, focusing on the eye region, which is consistent with the PCA analysis reported above[3].

[3] I also computed MLDA eigenvectiors using trial-level heatmaps, but given high intra-individual differences this produced a largely uninterpretable result, as shown in APPENDIX C.

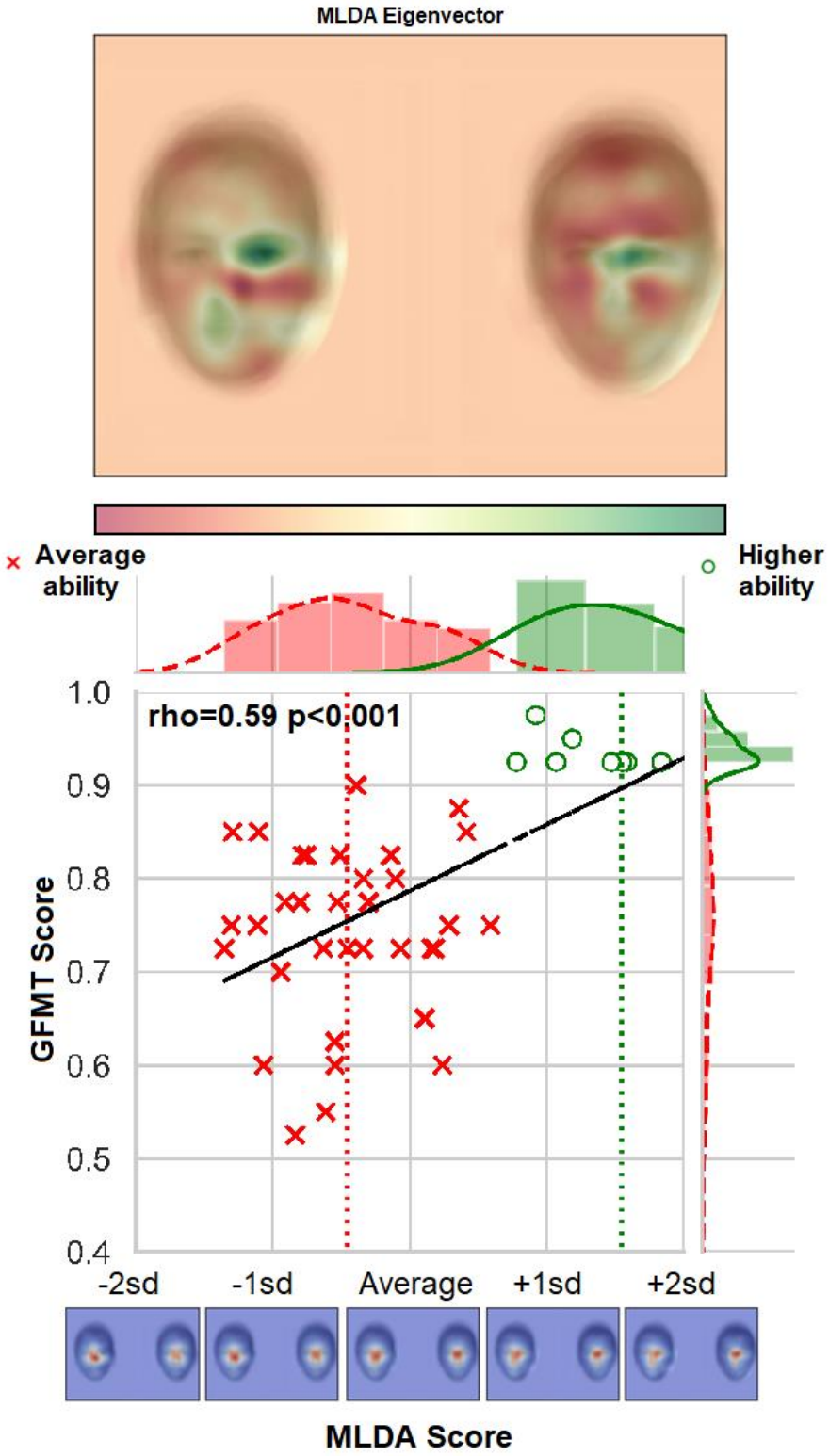


Figure 4.7. MLDA eigenvector and the correlation of MLDA scores with face-matching performance. At the top, we illustrate the MLDA eigenvector as a heatmap image showing the regions that best differentiate the fixation patterns between two participant groups. Participants in the Average ability group fixated more on the red areas of the depicted heatmap. In contrast, participants in the Higher ability group fixated more on the green areas. At the bottom, we show the MLDA Score distribution among participants, the average score for each group (vertical dashed lines), and the correlation between MLDA scores against face-matching ability.

To further investigate the MLDA scores distribution, we performed a correlational analysis of participants' MLDA scores against the five first PC loading scores. In this analysis, PC1 [rho(42)=0.59, p<0.001,CI = [0.3554 0.7547]] and PC3 [rho(42)=0.57, p<0.001, CI = [0.3287 0.7414]] loading scores significantly correlated with MLDA scores. This finding reiterates that the fixation pattern differences related to face-matching ability rely on the initial PCs (i.e. the most observable fixation pattern differences). We also investigated the relations between MLDA scores and ROI fixations and found that fixations on the right eyes [rho(42)=0.513, p<0.001, CI= [0.2549 0.7028]] significantly correlated with MLDA scores.

Individual difference analysis of visual exploration (Gini coefficient)

We investigated whether the amount of exploration participants engaged correlated with performance. Visual exploration was measured using the Gini coefficient (Lorenz, 1905), a widely used metric of data dispersal. Lower Gini coefficients represent greater data dispersal and, therefore, higher visual exploration in the present context. Table 4.3 shows correlations between participant-level heatmap Gini coefficients, face processing ability, 5 PCs, MLDA scores, and the average time spent performing the stimuli. In this section, we only report the participant-level data, but trial-level analysis shows similar results (see APPENDIX C).

First, Gini coefficients were significantly negatively correlated with face-matching ability, signalling that higher performance on the task was associated with greater dispersal of fixations (i.e. more visual exploration). Second, correlating the Gini coefficient scores with the PCA enabled us to examine which PCs were associated with greater exploration. PCs 3 and 4 showed relatively high, suggesting that these sources of inter-individual differences in fixation patterns might index individual differences in visual exploration. Third, greater exploration also correlated with the MLDA eigenvector, strengthening our conclusion that greater exploration is associated with higher accuracy on the GFMT. Fourth, we found high correlations between the Gini coefficient scores and the time that participants took to study face pairs in the test [rho(42)= -0.785, p<0.001, CI= [-0.8774 -0.6364]]. This result shows that those who studied the image pairs for longer also explored more facial regions, which may

explain why we also found that study time predicted accuracy on the task [rho(42)= 0.439, p=0.003, CI= [0.1634 0.6510]].

Correlation Matrix Gini coefficient

		Gini coefficient	
GFMT Score	Spearman's rho	-0.425	**
	p-value	0.004	
	CI	[-0.6410 -0.1466]	
PC1	Spearman's rho	-0.151	
	p-value	0.328	
	CI	[-0.4287 0.1527]	
PC2	Spearman's rho	-0.137	
	p-value	0.373	
	CI	[-0.4169 0.1667]	
PC3	Spearman's rho	-0.558	***
	p-value	< .001	
	CI	[-0.7334 -0.3130]	
PC4	Spearman's rho	-0.513	***
	p-value	< .001	
	CI	[-0.7028 -0.2549]	
PC5	Spearman's rho	0.118	
	p-value	0.444	
	CI	[-0.1854 0.4008]	
MLDA Score	Spearman's rho	-0.331	*
	p-value	0.029	
	CI	[-0.5717 -0.0378]	
Average Time elapsed	Spearman's rho	-0.785	***
	p-value	<.001	
	CI	[-0.8774 -0.6364]	

Note. * p < .05, ** p < .01, *** p < .001

Table 4.3. Correlation between participant-level Gini coefficients with GFMT Scores, Principal Component loading scores, MLDA scores, and time elapsed performing the stimuli.

Discussion

In Experiment 2, we investigated participants' fixation patterns when performing a standardised face-matching test (GFMT: Burton et al., 2010). We first delimited facial regions of interest (ROI) to investigate fixation patterns: the central area of the face (e.g. nose, eyes, and mouth) and one more prominent region that accounted for the remaining facial features of the face (e.g. peripheral areas such as ears, cheeks, etc.). We complemented this ROI analysis by examining the PCA of participant heatmaps, an exploratory technique that allowed us to illustrate the major sources of variance in heatmaps across participants, and MLDA, which compressed all the data into the single dimension that optimally discriminated high performers from participants with lower levels of accuracy. Cross-correlating PCA with regions-of-Interest data and a measure of visual exploration (Gini coefficient) enabled us to describe the major sources of inter-individual variance in fixation patterns and how this related to accuracy on the task.

This approach enabled us to explore our eye-tracking data, building a picture from converging sources of evidence to provide insight into the face information sampling strategies that might underpin individual differences in face identity processing ability. We found that higher performance on the GFMT was associated with two main differences in fixation patterns. First, greater attention to the eye region, particularly the right eye, as shown by both PCA (shown in PCA, MLDA and ROI analysis). Second, greater visual exploration of face information as evidenced by the Gini coefficient's association with GFMT accuracy, PC3 and the MLDA. The latter result appeared to be moderated by the time that participants spent studying the images, suggesting that more careful visual analysis that spreads attention across facial features is beneficial to unfamiliar face-matching accuracy.

ROI analysis showed that – on average - participants attended most of their fixations on the nose. Importantly, a greater focus on the nose region was not associated with accuracy. Given eye-tracking research suggesting that greater fixations on the nose might represent improved holistic processing (Bobak et al., 2017; Bennetts, Mole & Bate, 2017), and studies suggesting that holistic processing is associated with face recognition accuracy (Wang et al., 2012; Bobak et al., 2017), we expected a correlation between nose fixations

(i.e. central regions) and GFMT accuracy. However, these prior studies are mostly based on face recognition tasks. For unfamiliar face-matching tasks, like the GFMT, it has been suggested that a more piecemeal (i.e. feature-based) approach is associated with accuracy (Megreya & Burton, 2016; Towler et al., 2017).

Our Gini coefficient results are also consistent with a feature-based approach underpinning face-matching ability. Comparing and exploring more features appears to predict high face-matching accuracy, which is somewhat contrary to the notion that holistic processing drives face identity processing ability, at least for the pairwise face-matching task we studied here. The PCA and the MLDA eigenvectors showed fixation patterns covering larger facial areas were generally associated with higher accuracy on the task. Combined with the Gini coefficient, these results suggest that accuracy is proportional to the extent to which gaze patterns are distributed across the face images. We also found that distributed gaze patterns strongly related to the time participants spent performing the task, suggesting that high performers explored more facial areas for better decisions but at the cost of longer time processing trials.

Another salient result from the ROI analysis was that a large proportion (38.26%) of fixations fell outside the feature ROIs. This large value shows that the ROI approach is perhaps unable to capture the true complexity of gaze patterns in face-matching tasks. Similarly, the average heatmaps 'wash out' these potentially important sources of variance when generating average fixation patterns by combining data from individual trials. Notably, however, when we instead examined individual trials and the variability of gaze patterns here, we found very large intra-individual trial-to-trial variation in the distribution of fixations. This suggests that the role of visual exploration is an important component of performance on face-matching tasks that are typically overlooked in eye-tracking studies.

To examine whether these observations are also found in other types of face identity processing tasks, we apply the analysis techniques developed here to understand fixation patterns in participants completing a standard face recognition memory task in the next experiments. As discussed here, it is possible that the association between accuracy and visual exploration in Experiment 2 was due to the type of simultaneous face-matching task participants were engaging in. All previous studies demonstrating a link between holistic processing and face identity processing ability have used memory-based recognition tasks

(e.g. DeGutis et al., 2013; Wang et al., 2012), and so in Experiment 3, we tested whether the pattern of results we found here would be replicated in a face memory paradigm.

Experiments 3A & 3B

In Experiments 3A and 3B, we applied analysis approaches described in Experiment 2 to investigate participants' face information use during a face recognition memory test. Unlike face-matching tasks, where participants require no memory component, participants viewed faces in a learning phase where they were required to commit the faces to memory. Later, we showed these faces mixed with new faces, and participants had to decide whether or not they had viewed the face in the learning phase. Experiments 3A and 3B were part of a larger study (see Dunn et al., 2022), and my contribution to this study was to apply my exploratory PCA-based approach to the dataset to investigate whether fixation patterns can explain individual differences in face recognition ability. Here, I only present analyses led by me.

In addition to using a recognition memory paradigm in Experiment 3, there were two main differences from the previous experiment. First, we measured participants' face information sampling by restricting participants' viewing using a gaze-contingent aperture 'spotlight' paradigm (see Papinutto, Lao, Ramon, Caldara, & Miellet, 2017). Second, we examined individual differences in face identification ability by recruiting groups of 'super-recognisers' and comparing their performance and information use to typical viewers. Super-recognisers are people who show extremely high performance in face identity processing tasks, and they were selected in the following study from large-scale online testing.

Experiment 3A

In this experiment, participants viewed faces in both natural viewing conditions and also in conditions where faces were viewed through gaze-contingent apertures centred on their fixations. These apertures varied in size and enabled us to test the extent to which face learning and recognition in super-recognisers depends on being able to sample global face information on each fixation. This research question was the primary focus of the paper from which the dataset reported here was collected (see Dunn et al., 2022). We conducted

the analysis presented here to examine the differences in fixation patterns between super-recognisers and typical viewers. Although the effect of gaze-contingent aperture viewing on this was not of primary interest, it is included here for completeness.

Methods

Participants

A total of 72 participants participated in this study, 28 typical viewers and 44 super-recognisers. Super-recognisers were recruited based on their scores on three face identification tests before the study: (i) The Cambridge Face Memory Test Long-form (CFMT+: Russell et al., 2009); (ii) The Glasgow Face Matching Task (GFMT: Burton et al., 2010); (iii) The UNSW Face Test (Dunn et al., 2020). All super-recognisers achieved +1.7SD above the average score in each of these three tests. This criterion to select super-recognisers based on thresholds in several tests is strict (see Ramon, 2021). It has elicited groups with reliable outperformance of typical viewers in prior work (e.g. White, Wayne & Varela, 2022). For comparison, we recruited 28 'typical viewers' who signed up in return for AUD 20.00 via a research participation website targeting the general public hosted by the School of Psychology at UNSW Sydney.

After eye gaze data processing (see Apparatus and eye movement classification), we had to exclude the data from six super-recognisers and two typical viewers due to difficulties with the eye-tracker (N=2), incomplete data (N=3), and data loss (+20% of trials removed, N=3). In total, we analysed data from 26 typical viewers (15 females, ten males, and one non-binary) aged between 18 and 61 years old (M=27, SD=8.6) and 34 super-recognisers (14 females, 20 males) aged between 26 and 51 years old (M=37.8, SD=7.4).

Stimuli

Stimuli were images of 144 faces identities of different genders, ethnicities, ages, and facial expressions collected from the Lifespan Database of Adult Facial Images (Minear & Park, 2004). To ensure the faces showed the same size to participants, we aligned the faces by the eye and mouth. Using the eye tracker (see below), we created several spotlight aperture viewing conditions using the method described by Papinutto and colleagues

(2017), with each face viewed through one of five aperture sizes (5º, 10º, 15º, 20º, and 25º of visual angle) or a natural viewing condition (i.e. no aperture size). See APPENDIX C for more detailed information about the chosen aperture sizes.

Apparatus and eye movement classification

We recorded participants' eye gaze data with Tobii Pro Spectrum. Participants sat in front of the eye tracking device and used a chin rest so that the angular size of the stimuli was 16.5° of visual angle. This device has an average gaze position error of about 0.25° and a spatial resolution of 0.01°. Before data collection, we conducted a nine-point fixation calibration procedure and repeated the process until reaching the optimal criterion. We only tracked participants' dominant eye. We classified eye gaze data using the method described in Experiment 2.

Procedure

After calibration, participants completed the face memory test. Faces were in six blocks, each containing a learning and recognition phase. In the learning phase, participants viewed 12 faces for 5 seconds each in one of the viewing conditions (five aperture sizes or Natural View, see Stimuli). A fixation cross appeared before each face's presentation to cue the screen centre, and faces were presented in random positions on the screen to avoid biasing their first fixations. The recognition phase immediately followed the learning phase, displaying a series of 24 faces (12 identities previously shown and 12 unfamiliar faces). Participants had to indicate whether the presented face was from a previously shown face via key press. In addition, participants that were typical viewers completed the CFMT+ (Russell et al., 2009) after the stimuli, whereas super-recognisers had already completed face tests as part of pre-screening (see Participants).

Fixation Analysis: Heatmaps

We generated heatmaps and conducted a Principal Component Analysis (PCA) of participants' average heatmaps using the same method as in Experiment 2. We used only the trial-level heatmaps in these analyses because of the introduction of stimulus viewing

conditions in Experiment 3. We first resized the original heatmap images from *891x656* to *209x182* pixels and converted each to a single dimension vector (i.e. one heatmap = 1x38038 pixels). We reduced the image size for less computational effort. Then, we normalised these vectors and applied the PCA on the transformed *Nxn matrix,* where *N* represents the number of heatmaps, and *n*, the resolution. We visualised these PCs as their dimension interacting with the average heatmap. We calculated the PCA for the learning and recognition phases separately. Because our data comprises two well-defined groups, we do not report correlation values for this experiment. Instead, we calculated linear mixed models to note differences between groups using the trial-level heatmaps.

Exploration analysis: Gini coefficient

As in Experiment 2, we measured the overall dispersion of participants' fixation patterns using the Gini coefficient (Lorenz, 1905). As a reminder, higher Gini coefficients represent more concentrated fixation patterns, and lower Gini coefficients represent more dispersed patterns.

Results

Overall accuracy

Figure 4.7 shows the accuracy (A-prime; see Stanislaw & Todorov, 1992) of super-recognisers and typical viewers across the viewing conditions. Although the viewing condition manipulation was not critical for this investigation, it is an important context for understanding the following analysis, so it is included here. Visual inspection of Figure 4.7 suggests that super-recognisers were superior in all viewing conditions. To test this assumption, we ran a linear mixed model. For the model of accuracy, we set participants' intercept as a random effect and group and aperture size as fixed effects. We found a significant main effect of group [b = 0.069, CI = [0.035, 0.102], t(58) = 4.05, 249 p < .001] and a non-significant interaction with aperture size [b= 0.017, CI=[-0.005,0.039], t(298)=1.52, p=0.129], showing that super-recognisers outperformed typical viewers irrespective of viewing condition. There was also a significant main effect of aperture size [b = 0.123, CI =

[0.112, 0.134], t(298) = 21.74, p < .001], whereby performance was generally poorer for smaller apertures.

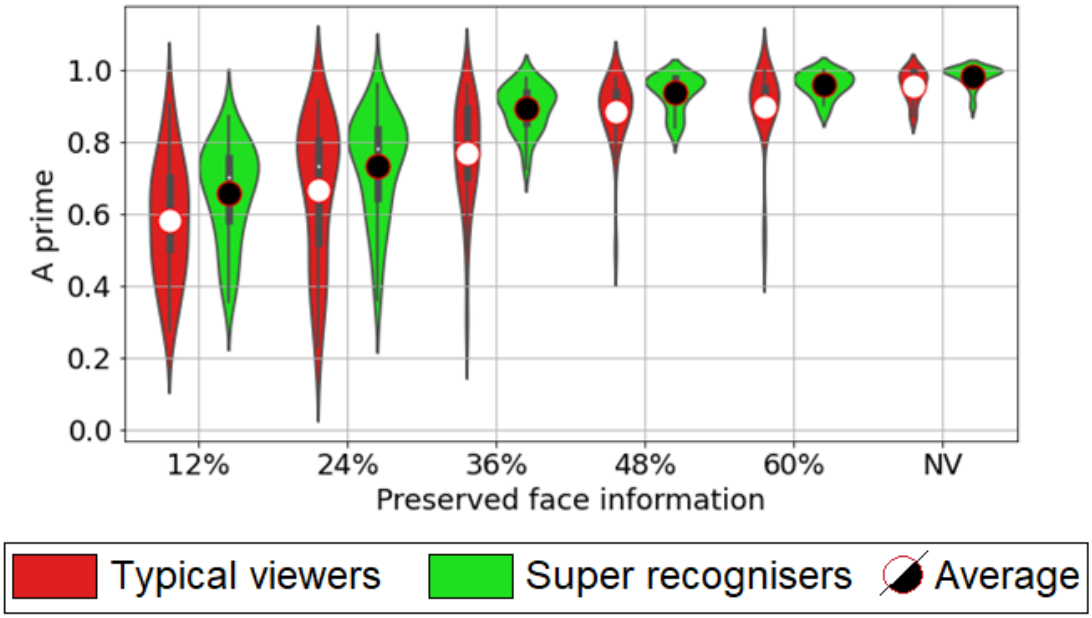


Figure 4.8. Results of the obtained accuracy (A prime) for typical viewers and super-recognisers across viewing conditions. We show the violin distributions for typical viewers in red and their respective averages in white; Super-recognisers' distributions are in green and their respective averages in black. The boxes show interquartile range, and whiskers delimit the max and minimum for each distribution.

Principal Component Analysis

We performed separate PCAs for the learning and recognition phases to investigate the main source of variation in fixation patterns across participants. We conducted this at the trial level, using heatmaps from individual trials. We show the visual reconstruction of the first 5 PCs for the learning and recognition phases in Figure 4.9. As a reminder, the average heatmap is represented with a loading score of zero for each PC, and a zero-mean normal distribution represents participants' loading scores across each component. Therefore, for every PC, some participants received a negative loading score (i.e. to the left side of the average), and some received a positive loading score (i.e. to the right side of the average).

We divided Figure 4.9 into two separate panels to show the two distinct phases of the experiment. On the left of Figure 4.9, we plot the visual reconstruction of the 5 PCs from

the learning phase of the experiment and the recognition phase on the right. Importantly, we calculated each panel using different heatmaps - derived from each phase. Still, a visual inspection of Figure 4.9 reveals striking similarities between both panels. This consistency of PCs between phases could indicate that participants use similar approaches to sample information when learning or recognising faces. However, a visual inspection might not suffice such an assumption. To test this, we repeated the procedure reported in experiment 2, using participants' individual; heatmaps projected on the respective components as PC scores to investigate individual differences between groups of participants (super-recognisers and typical viewers).

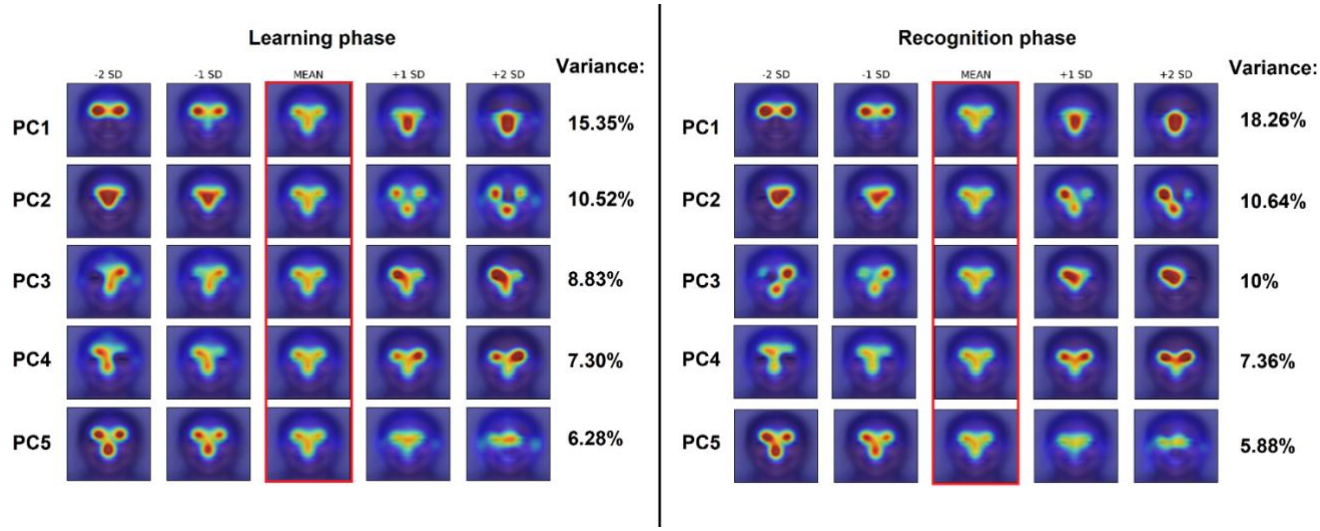


Figure 4.9. PCA navigation for Learning (Left) and Recognition (Right) phases using all aperture conditions of the study. We show the five first Principal Components and their respective explained variance.

We investigated whether PC1 could differentiate super recognisers from typical viewers by calculating PC1 loading scores for each group separately for the learning and recognition phases (see Figure 4.10). In both phases, PC 1 appears to differentiate participants from sampling more information from the eye region of the face to engaging in more central areas of the face, which is consistent with the trial-level PCA conducted in Experiment 2 (see Figure 4.6). In Figure 4.10A, we replicate the PC1 navigation illustrated above. The PC1 for both phases shows that some participants tended to fixate on the eye region of faces, receiving a negative PC1 loading score. In contrast, participants who received positive loading scores broadly distributed their fixations across the central facial

region. In Figure 4.10B, we show the score distribution between participant groups for the two phases of the experiment.

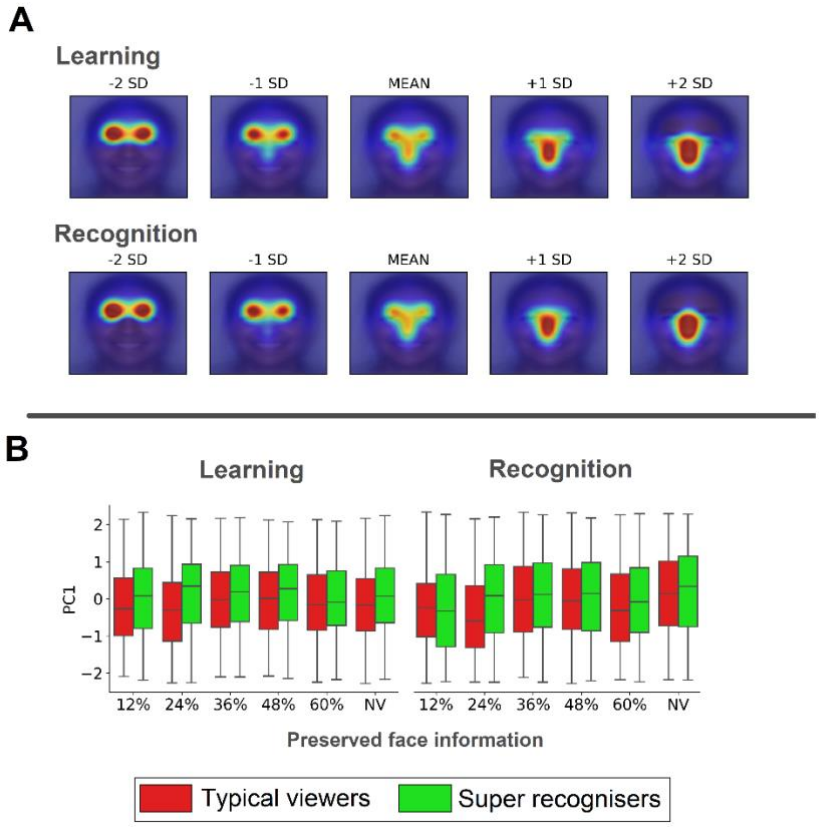


Figure 4.10. Analysis of PC1 for the trial-level fixation heatmaps during face learning and recognition phases. (A) The interaction of the average fixation heatmap with calculated PC1 components for both phases. (B) Boxplots showing PC1 loading scores across aperture conditions. The boxes show the interquartile range, and the whiskers show the minimum and maximum of the distribution.

We conducted a linear mixed model analysis with participants' intercept as a random effect, and group, aperture size, and phase as fixed effects. Although aperture size had a significant effect on PC1 scores, the three-way interaction of group, phase and aperture condition was non-significant [b=-0.01, CI =[-0.07, 0.04], t(11516.6)= -0.46, p= 0.642]. Importantly, because the aperture viewing conditions were not of primary interest for the line of investigation in this thesis, I next conducted a separate analysis using only the study's natural viewing condition (i.e. non-aperture). The analysis using only the natural viewing condition will be more beneficial in examining whether patterns observed in Experiment 2 were similar in this recognition memory task. The full 3-factor LMM is available in Appendix C – Experiment 3A.

Principal Component Analysis (Natural Viewing condition only)

Separate PCAs were conducted for the learning and recognition phases using the gaze heatmaps derived from the Natural Viewing condition of the study. Visualisations of PC1 for natural viewing conditions only are shown in Figure 4.11. Visually comparing Fig 4.11 to Figure 4.9, we observe some commonalities between the generated PCs for the aperture condition and natural viewing. For example, the first component (PC1). PC1 represents differences in participants' tendency to fixate on the eye region of faces versus more broadly distributing fixations across the central facial region. For the learning phase, PC1 accounted for ~15% of the total variance and ~22% for the recognition phase.

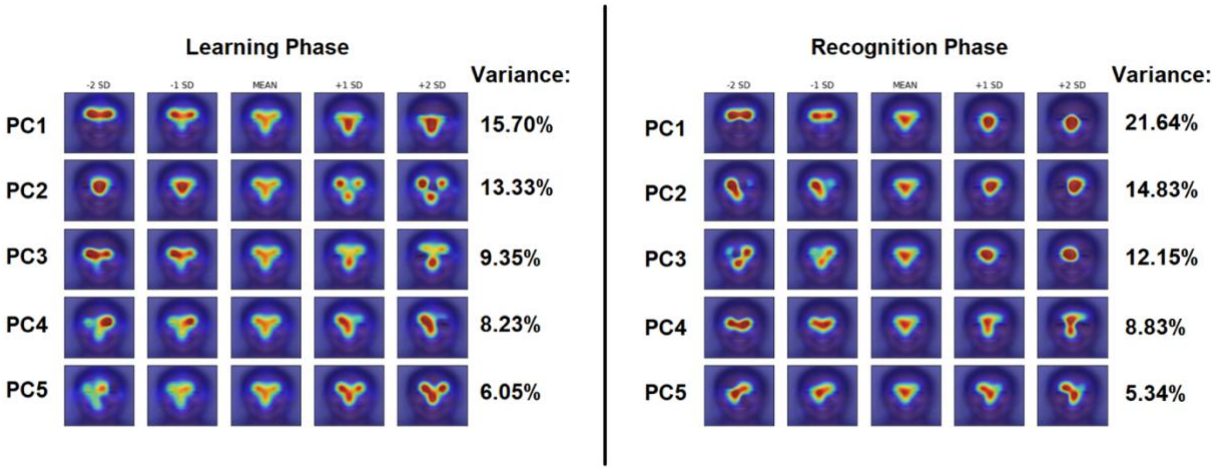


Figure 4.11. PCA navigation for Learning (Left) and Recognition (Right) phases using only the NV aperture condition of the study.

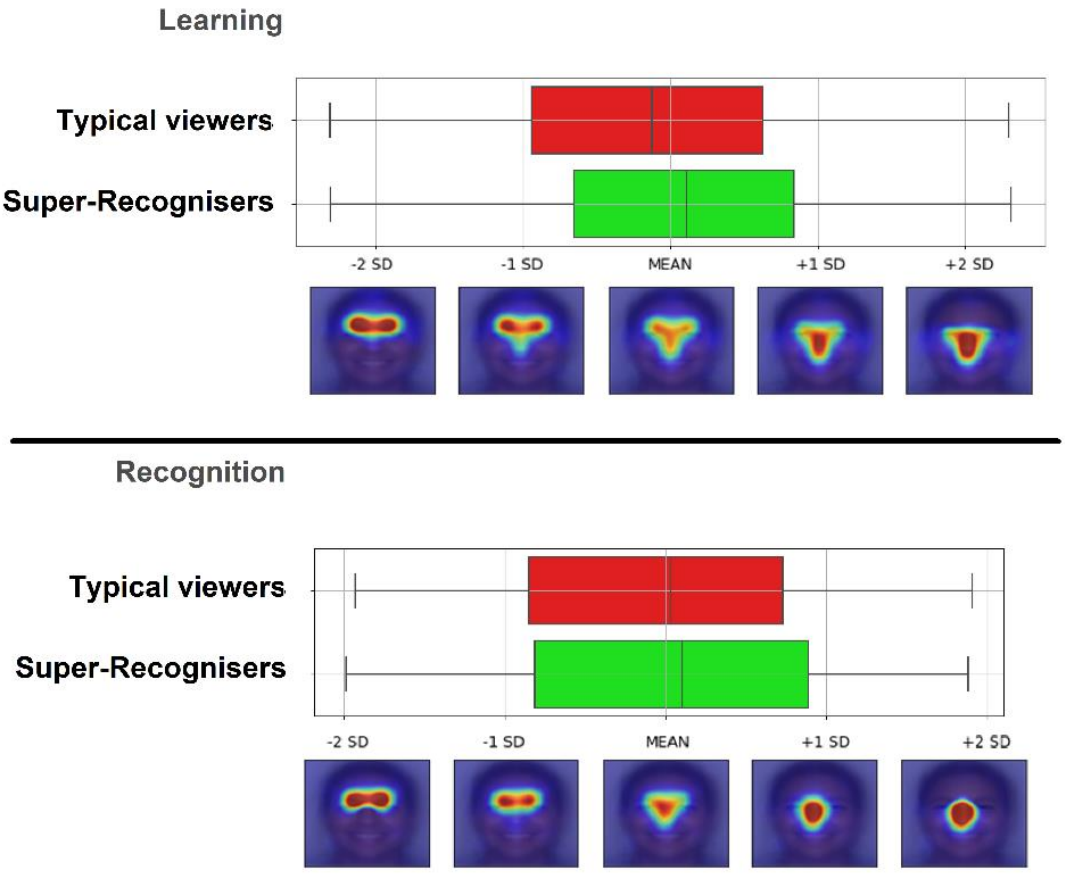


Figure 4.12. Analysis of PC1 for the trial-level fixation heatmaps during face learning (Top) and recognition (Bottom) phases considering only the Natural Viewing (NV) aperture condition. We show the interaction of the average fixation heatmap with calculated PC1 components for the two separate phases and boxplots, showing the range of PC1 loading scores for the NV aperture condition. The boxes show the interquartile range, and the whiskers show the minimum and maximum of the distribution.

We used linear mixed model analysis to analyse the pattern of results in Figure 4.12, with participants' intercept as a random effect and group (typical viewers, super-recognisers) and study phase (learning, recognition) as fixed effects. The main effects of group [b = 0.09, CI = [-0.28, 0.47], t(58.1) = 0.492, p = 0.625], and phase [b = -0.3, CI = [-0.09, 0.04], t(1745.9) = -0.854, p = .393] were non-significant. However, we observed a significant interaction between phase and group [b = -0.3, CI = [-0.43, -0.17], t(1745.9) = -4.47, p < .001]. We visualised this interaction in Figure 4.13 for clarity and analysed simple main effects, which showed a larger difference between groups during face learning [b=0.245] than recognition [b=-0.057]. Post-hoc comparisons showed super-Recognisers had more positive PC1 values than typical viewers during learning (t(1802)=3.22, $p_{bonferroni}$=0.008) but

not during recognition phase (t(1802)= 0.628, $p_{bonferroni}$=1). This result shows that Super-recognisers explored the face more than Typical Viewers during face learning but not during recognition. Therefore the tendency we observed in Experiment 2 for high performers to explore the face more in face-matching tests appears also to be true of high performers in recognition memory tests, but only when they initially learn the faces.

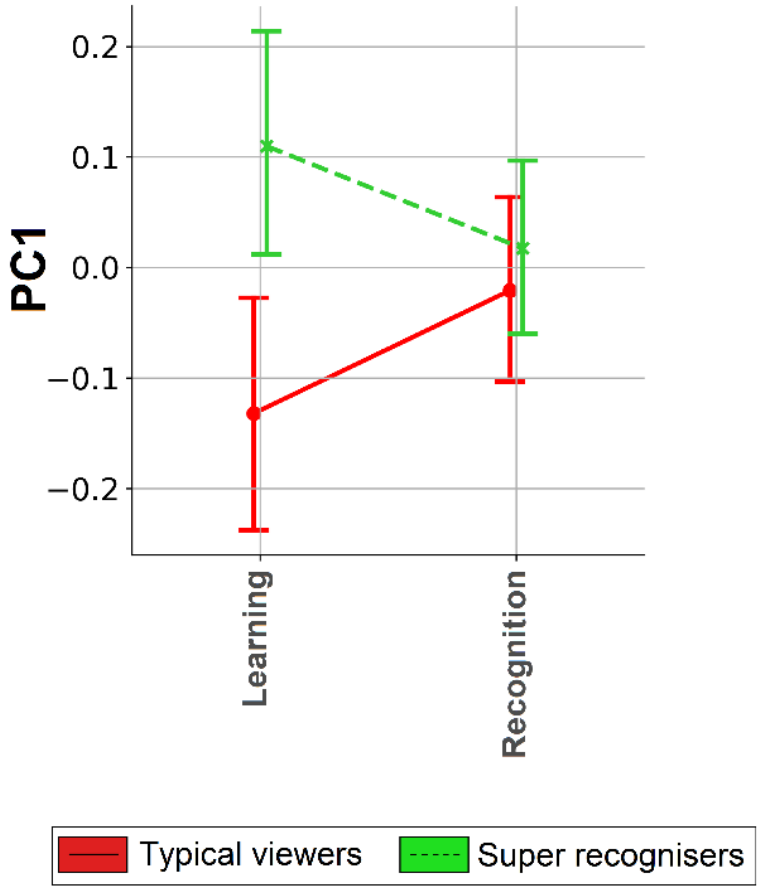


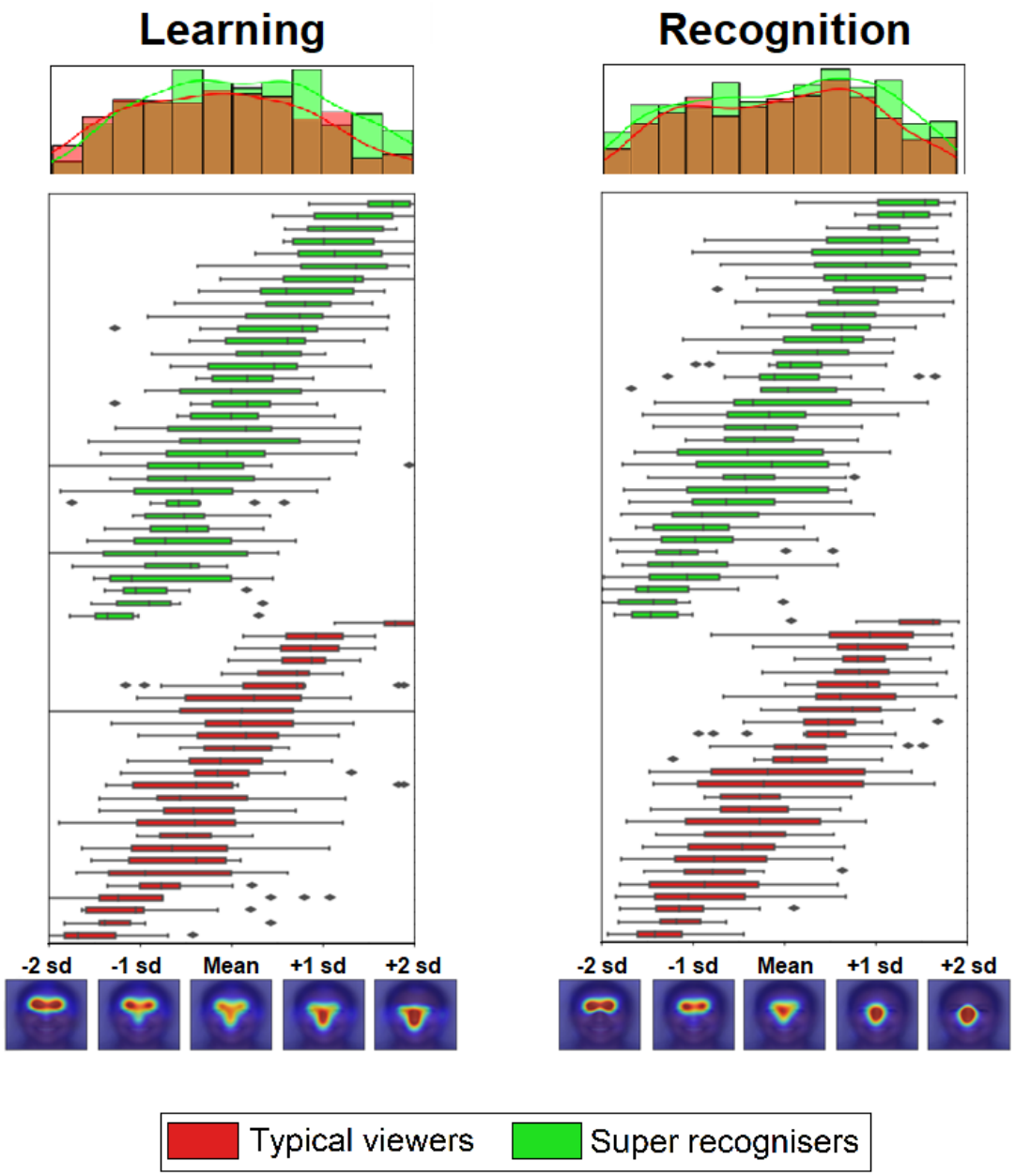
Figure 4.13. Categorical plot showing the interaction of PC1 scores for the two groups of participants in the two phases of the study during the Natural Viewing aperture condition. Whiskers show a 95% confidence interval.

Intra-individual difference analysis of fixation patterns (Natural Viewing condition only)

The previous analysis categorised the relationship between PC1 scores and face processing ability. This analysis found that the most significant difference between super-recognisers and typical viewers, measured by the first component, is when learning faces. However, as we discovered in Experiment 2, it is also important to measure the variability of fixation patterns (PC1 scores) for individual participants across the different trials.

Figure 4.14 shows the inter-trial variability of PC1 scores across all participants processing the Natural Viewing condition of the experiment. On the left panel, we show the results for the learning phase of the experiment, and on the right panel, the recognition phase. We also show results for typical viewers in red and super-recognisers in green. As with Experiment 2, a visual inspection of Figure 4.14 reveals an extensive range of PC1 scores for the two groups processing the stimulus trials, suggesting again that participants were flexible in their strategies to sample information from faces. To again found evidence of wide variability in PC1 scores across trials, with an average range of participants' PC1 scores (*range = max - min*) in the learning phase of 2.17 standard deviations (Typical viewers= 2.19; Super-recognisers= 2.15; t(58)=0.217, p=.829), and in the recognition phase 2.17 standard deviations (Typical viewers= 2.13; Super-recognisers= 2.20; t(58)=0.56, p=.578)[4].

[4] We found a similar range for other PCs. For a similar analysis but considering all aperture conditions, see APPENDIX C - Experiment 3A.



4.14. Trial-level variability in PC1 loading scores across participants for the experiment's learning (left) and recognition (right) phases. Consistent with previous graphs, we show results for typical viewers in red and super-recognisers in green. This data only considers the natural viewing condition of the experiment. We ranked PC scores by their mean value for each panel separately, so a side-by-side comparison does not measure the same participant. The boxes show the interquartile range, and the whiskers show the minimum and maximum of the distribution.

Gaze dispersal Analysis (Natural Viewing Only)

The previous analysis indicated that super recognisers explored faces more during learning. However, because we based our understanding of PC1 on our visual interpretation

of the component, we conducted a more direct visual exploration by calculating the Gini coefficient for each trial-level heatmap (see Methods). As a reminder, lower Gini Coefficients denote greater gaze dispersal, an indication of more visual exploration.

To make the analysis of Gini coefficients consistent with the PCA analysis - and to allow us to compare results with Experiment 2 - we only used the Natural viewing aperture condition of the study (for all aperture conditions, see APPENDIX C – Experiment 3A). We used a linear mixed model setting participants' intercept as a random effect and group and phase as fixed effects, which revealed a significant main effect of phase [b=0.049, CI= [0.044, 0.054], t(1562)= 18.438, p< 0.001] and a non-significant main effect of group [b=-0.004, CI= [-0.02, 0.012], t(43.8)= -0.49, p= 0.626].

These main effects were qualified by a significant two-way interaction between group and phase [b= 0.025, CI= [0.014, 0.035], t(1562)= 4.67, p< 0.001]. Investigating this interaction further, simple effects reveal a larger difference between super recognisers and typical viewers during the learning phase [b=-0.016] than in the recognition phase [b=0.008]. Consistent with the PCA analysis, this shows that the differences in viewing behaviour between super-recognisers and typical viewers were most pronounced during face learning, with super-recognisers showing more visual exploration.

Discussion

In Experiment 3A, we investigated the fixation patterns of typical viewers and super-recognisers processing a face-recognition task by applying the previously introduced PCA technique. We used the PCA to reduce our data to reveal core aspects of the nature of fixation patterns amongst our participant groups. By subdividing fixation patterns into principal components, we found that the first components significantly explained some differences between super-recognisers and typical viewers. This result is interesting because it shows that individual differences in face recognition ability could be related to major sources of variation in how people sample information from faces. This major source of variation (PC1) was that some participants focused more on the eye region while others distributed their gaze more widely, which was consistent with the trial-level analysis in Experiment 2.

Interestingly, the extent to which PC1 discriminated between typical viewers and super-recognisers depended on the phase of the experiment, with differences being observed more prominently during face learning. This result illustrates that super-recognisers investigate more facial regions to extract more face information when learning new faces. The PCA using only the Natural View condition of the study revealed similar results compared to the complete data using all aperture viewing conditions, suggesting that the tendency to explore the face is also present when the entire face is in view. Analysis of the Gini coefficient strengthened this conclusion, with super recognisers showing greater dispersion throughout facial features during face learning.

Relating this to the results of Experiment 2, we have accumulated evidence for the importance of visual exploration beyond the central and eye regions of the face in achieving high accuracy in face identity processing tasks. This was observed both when matching faces showed simultaneously on the screen (Experiment 2) and when studying faces in order to commit them to memory (Learning phase, Experiment 3). This perhaps suggests that high performers and super-recognisers were more efficient in extracting face identity information, allowing them to extract more elaborate identity information.

Therefore, understanding the differences between typical viewers and super-recognisers on how they first encode face information might help us understand how super-recognisers develop more robust memory representations of faces. Super-recognisers showed an advantage in processing the stimuli even when the amount of information available per fixation was minimal. This result is important because studies aiming to understand individual differences in face processing ability initially associated measures of holistic processing with face identification performance (DeGutis, Wilmer, Mercado, & Cohan, 2013). This association is still debatable because some studies reported evidence of enhanced holistic processing predicting individual differences (e.g. Wang et al., 2012), while recent studies have not (e.g. Sunday et al., 2017). Nevertheless, our results corroborate – and extend – the suggestion of a more local/featural processing in enhanced face processing ability (Royer et al., 2015; Tsantani et al., 2020).

These results also broaden the understanding of information used to identify faces (e.g. Chuk, Chan, & Hsiao, 2017; Royer et al., 2018; Schyns, Bonnar & Gosselin, 2002; Tardif et al., 2019). We show that single virtual components created by an orthogonal linear transformation (i.e. PCA) can identify differences between typical viewers and super-

recognisers. Exploring these components further, our analysis of PC1 revealed that super-recognisers fixated less on the eye region than typical viewers. Curiously, previous studies report the eye region as crucial for recognition (e.g. Bate, Haslam, Tree & Hodgson, 2008; Schyns, Bonnar & Gosselin, 2002; Slessor, Riby & Finnerty, 2013) and avoidance of the eye region often relates to poorer face recognition ability and even prosopagnosia (e.g. Caldara et al., 2005; Lê, Raufaste, Roussel, Puel & Démonet, 2003; Lê, Raufaste & Démonet, 2003; Stephan & Caine, 2009; Towler et al., 2016; Avidan & Behrmann, 2021). In addition, previous work of - for example - Tardiff and colleagues (2019) suggests that super recognisers processing is specially tuned to information in the eye region, at least when the presentation of faces was restricted by 'bubbles' apertures used in that study (see Schyns et al., 2002).

It is still a topic of debate in the face-processing literature if the superiority of super-recognisers is because they are an extreme version of a typical viewer or because their mechanisms of face-processing are completely distinct from typical viewers (Noyes et al., 2017; Tardif et al., 2019). In other words, it is still unclear if the differences between super recognisers and typical viewers are quantitatively or qualitatively different. In Experiment 3A, we compared the fixation patterns of 26 typical viewers and 34 super-recognisers. And so, due to the majority of super-recognisers used in this study, it raises the question if the primary components detected differences between the two groups because they are part of a continuum of processing differences or because they use qualitatively different processes. In Experiment 3B, we aim to answer this question by testing whether the dimensions explaining variance in typical viewers viewing behaviour can generalise to capture what super-recognisers are doing.

Experiment 3B

In Experiment 3A, super-recognisers and typical viewers used different fixation patterns to learn faces. However, it is still unclear whether these differences reflected the qualitative differences between the two well-defined groups or whether they reflect underlying continuous variation across the face identity ability spectrum. In Experiment 3B, we aimed to address this question by recruiting a larger group of typical viewers - in relation

to super-recognisers - to examine whether the dimensions explaining their eye strategy were still similar when compared to the ones found in Experiment 3A.

Methods

Participants

We recruited 43 typical viewers from UNSW Sydney and the University of Wollongong. Analysis was based on 42 typical viewers (35 females, 7 males, M_{age} = 21.2, SD_{age} = 5.0) and 3 super-recognisers, with one typical viewer excluded due to excessive data loss. All participants had normal or corrected to normal vision.

Stimuli

We used the same Stimuli used in Experiment 3A.

Apparatus and eye movement classification

We recorded participants' eye movements using the SR-Research EyeLink Portable Duo and classified fixations the same way as Experiments 2 and 3A.

Results

Overall Accuracy

We show the relation of accuracy (A prime scores) across all aperture size conditions in Figure 4.14. Visual inspection suggests a linear trend resulting in better accuracy for larger aperture sizes. To test this, we ran a linear mixed model. For the accuracy model, we set participants' intercept as a random effect and CFMT+ and aperture size as fixed effects. We found a significant main effect of CFMT+ for A prime [b= 0.042, CI= [0.023, 0.060], t(43)= 4.42, p< .001] and a significant main effect of aperture size [b= 0.091, CI= [0.08, 0.101], t(223)= 16.61, p< .001]. Consistent with Experiment 3A, the interaction of CFMT+ and Aperture size [b = -0.004, CI= [-0.015, 0.007], t(223)= -0.69, p = 0.492] was not significant.

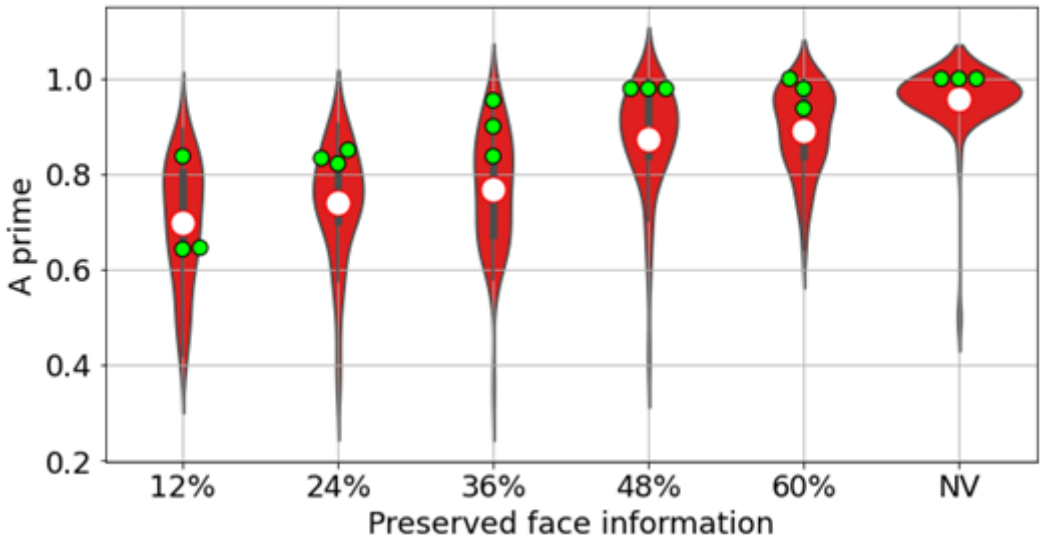


Figure 4.14. Results of the obtained accuracy (A prime) for typical viewers and super-recognisers across viewing conditions. We show the violin distributions for typical viewers in red and their respective averages in white. And show individual super-recognisers in green.

Principal Components Analysis

As in Experiment 3A, we performed separate PCAs for the learning and recognition phases to investigate fixation patterns between participants. Similar to the previous analysis, we calculated two distinct PCAs, considering each phase's trial-level data (intra-individual variation).

The first five PCs emerging from the PCA of typical viewer data are shown in Figure 4.15. Visually inspecting Figure 4.15, PCs calculated for the learning and recognition phases are very similar, and they are also very similar to components found in Experiment 3A. This visual similarity is important because it suggests that super recognisers do not differ qualitatively from typical viewers in sampling information from faces. Rather, their gaze behaviour is captured by components of variance that are present across typical viewers, pointing to the continuity of processing differences across the ability spectrum rather than distinct processing mechanisms in super-recognisers compared to typical viewers.

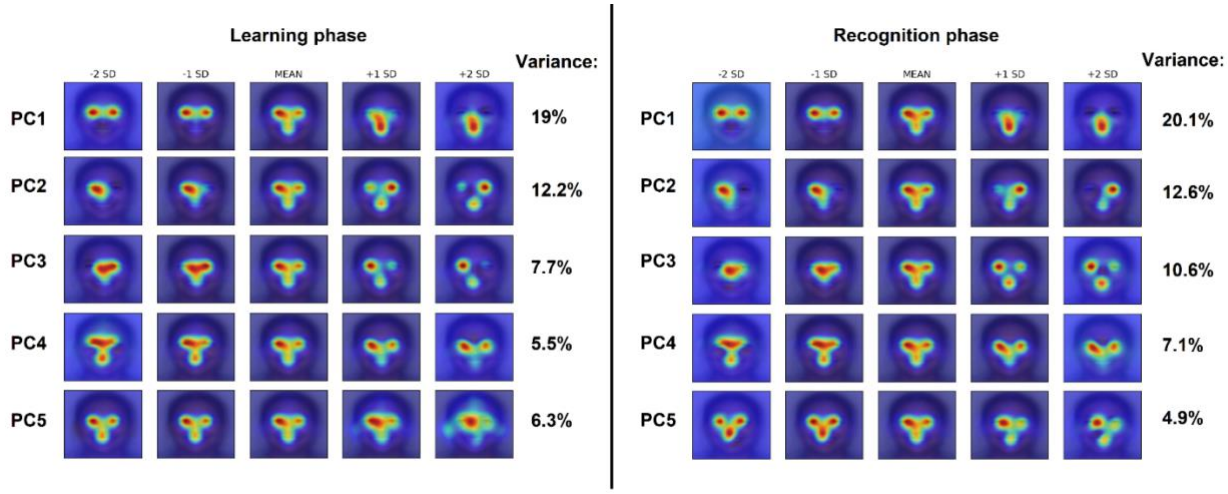


Figure 4.15. PCA navigation for Learning (Left) and Recognition (Right) phases using data from all aperture conditions.

We conducted further analysis of PCA loadings on the PCA in Figure 4.15 which was constructed using data from all aperture viewing conditions. However, as in Experiment 3A, the results of this analysis were similar to the analysis considering only the Natural Viewing condition. For simplicity and for consistency with Experiments 2 and 3A, we therefore only present analysis for the Natural Viewing condition below. For the full analysis, please refer to APPENDIX C – Experiment 3B.

Principal Components Analysis (Natural Viewing only)

We replicated the PCA from the previous section only considering the experiment's Natural Viewing (NV) condition. We plot the reconstruction of the first five principal components in Figure 4.16. Interestingly, a visual inspection of Figure 4.16 shows striking similarities with the previously calculated in Figure 4.15. This consistency suggests that the strategies to process faces remained comparable across aperture viewing conditions.

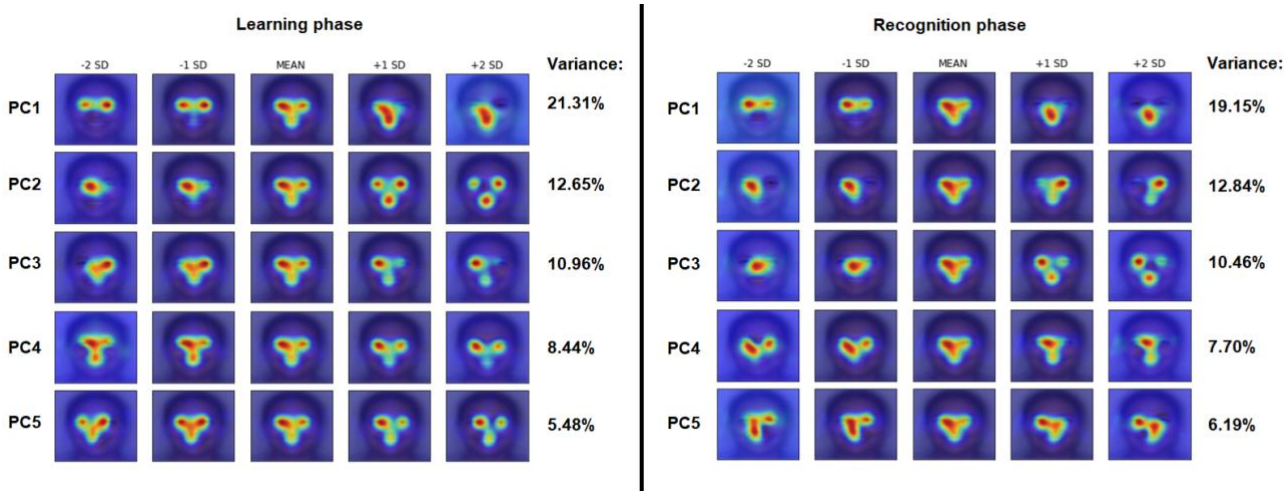


Figure 4.16. PCA navigation for Learning (Left) and Recognition (Right) phases using data only from the Natural View (NV) aperture condition.

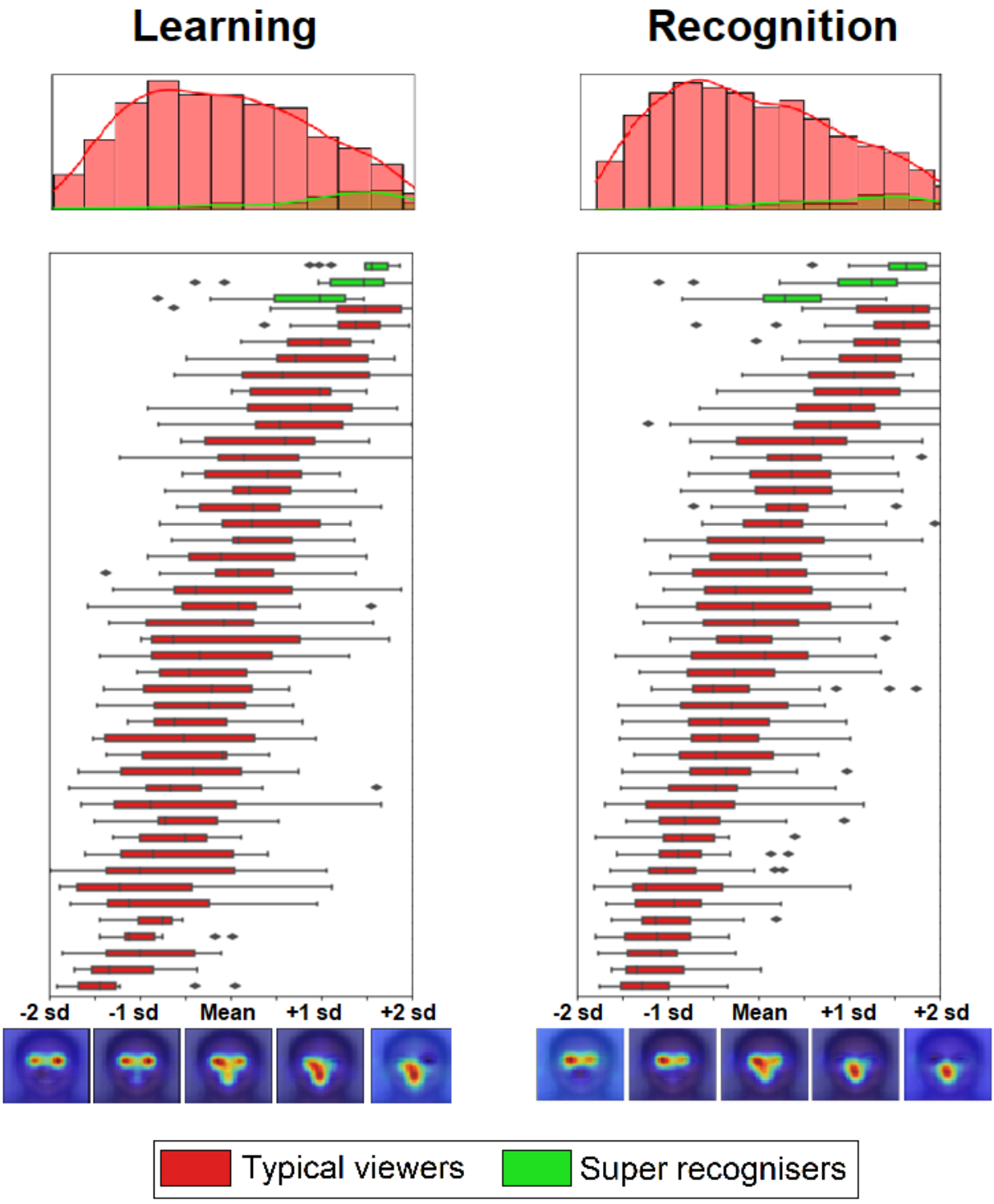
First, we used linear mixed model analysis to analyse the pattern of results for PC1 scores. For that, we set participants' intercept as a random effect and CMFT+ scores and phase as fixed effects. We found that both main effects of phase [b = 0.006, CI = [-0.065, 0.077], t(1562) = 0.17, p = 0.863] and CFMT+ [b = 0.006, CI = [-0.011, 0.022], t(43.2) = 0.65, p = 0.519] were not significant. In addition, we observed a non-significant interaction between phases and group [b = -0.003, CI = [-0.008, -0.002], t(1562) = -1.19, p = 0.235]. This non-significant result shows that PC1 loading scores could not differentiate typical viewers based on their CFMT+ score. Still, in this analysis, the simple effects -using phase as a moderator- illustrate a larger difference during face learning [b=0.007] than recognition [b=0.003], showing higher PC1 scores for those with higher CFMT+ scores. And so, assuming that the two separate PC1 calculate the same component, these results show that, among typical viewers, those with higher scores tended to sample information from faces more during face learning avoiding the eye region, despite simple effects not being significant. We describe the model for other PCs in APPENDIX C – Experiment 3B.

Intra-individual difference analysis of fixation patterns (Natural Viewing only)

Similar to Experiment 3A, the previous analysis categorised the relationship between PC1 scores and face processing ability. This analysis showed no traces that high performers processed faces differently than low performers, measured by the first component. However, it is essential to address that participants of both groups could have received a

wide range of PC scores in the component. Therefore, to investigate stability in fixation patterns across participants (i.e. PC1 loading scores), we show their inter-trial differences in Figure 4.17.

Figure 4.17 shows the inter-trial variability of PC1 scores across all participants processing the Natural Viewing condition of the experiment. On the left panel, we show the results for the learning phase of the experiment, and on the right panel, the recognition phase. We also show results for typical viewers in red and super-recognisers in green. However, in this analysis, our group of 3 super recognisers are coloured just for illustrative purposes. A visual inspection of Figure 4.17 reveals an extensive range of PC1 scores for participants processing the stimulus trials. And so, similar to Experiment 3A, this range illustrates that participants were flexible in their strategies to sample information from faces. To illustrate the order of magnitude of this flexibility (in PC1 scores), we found that the average range of participants' PC1 scores (*range = max - min*) in the learning phase was of 2.27 standard deviations, and in the recognition phase was of 2.35 standard deviations. And so, this vast range of PC1 scores per participant possibly illustrates that individual trials influence participants to engage in particular fixation patterns for the NV condition. We found a similar range for other PCs.



4.17. Trial-level variability in PC1 loading scores across participants for the experiment's learning (left) and recognition (right) phases. Consistent with previous graphs, we show results for typical viewers in red and super-recognisers in green. This data only considers the natural viewing condition of the experiment. We ranked PC scores by their mean value for each panel separately, so a side-by-side comparison does not measure the same participant necessarily. The boxes show the interquartile range, and the whiskers show the minimum and maximum of the distribution.

Gaze dispersal Analysis (Natural Viewing only)

In Experiment 3A, we found significant differences between groups of participants when analysing their dispersal when learning and recognising faces. For that, we calculated the Gini coefficients for every heatmap generated by participants' fixations (see Methods) as a metric of exploration. As a reminder, a lower Gini Coefficient would represent higher fixation activity (i.e. higher exploration). Here, we replicated the analysis considering the natural viewing aperture condition of Experiment 3B. For all aperture conditions, see APPENDIX C – Experiment 3B.

We used a linear mixed model setting participants' intercept as a random effect and group and phase as fixed effects. Linear mixed models reveal a significant main effect of phase [b=0.04, CI= [0.038, 0.044], t(1409.7)= 28.51, p< 0.001] and a not significant main effect of CFMT+ [b= -1.73e-4, CI= [-4.81e-4, 1.34e-4], t(43.8)= -1.10, p= 0.276]. However, we found that the two-way interaction between group and phase [b= 4.48e-4, CI= [9.82e-5, 6.41e-4], t(1537.8)= 4.57, p< 0.001] was significant. Investigating this interaction further, simple effects of CFMT+ reveal a significant effect during the learning phase [b=-3.97e-4] and a not significant effect in the recognition phase [b=5.10e-5]. This analysis shows that participants explore more facial regions (i.e. lower Gini coefficient) when learning faces, and those with higher face processing ability explore significantly more.

Discussion

In Experiment 3B, we used a larger cohort of typical viewers to find similar results compared to experiment 3A. In experiment 3A, we used groups of super recognisers and typical viewers processing a face recognition task. Our objective with that study was to investigate different fixation patterns between groups by analysing fixation heatmaps. Using principal components, we found that the main sources of variation in these heatmaps successfully measured differences between groups of participants. However, it was still unclear if the calculated PCs reflected two distinct groups of participants with different fixation patterns or a continuum. Therefore, in Experiment 3B, we investigated a group of typical viewers to distinguish if super-recognisers could be an extreme version of typical viewers.

Measuring individual differences using the continuous variable of CFMT+ score, we found remarkable consistency with the results of previous experiments using binary super-recogniser / typical viewer group membership. This consistency of results is significant because, first, it shows robustness in the linear trends seen by the PCA when investigating participants across different face processing ability spectrums. Second, it suggests that super-recognisers' mechanisms to process faces are not necessarily qualitatively different from typical viewers (see Noyes et al., 2017) because we generated similar principal components using typical viewers. And so, Experiment 3B widened previous studies which examined information use by high-performers in face identification tasks (Chuk, Chan, & Hsiao, 2017; Royer et al., 2018; Schyns, Bonnar & Gosselin, 2002; Tardif et al., 2019). That is, high-performers tend to sample more information when learning faces, and the primary source of variation suggests that they avoid the eye region when doing so.

Experiment 4

Experiments 2 and 3 showed significant differences in fixation patterns between participants across the face-processing ability spectrum. In these experiments, we found that participants at the upper end of the face-processing ability spectrum showed enhanced face exploration compared to typical viewers. This shows that super-recognisers sampled more face information than typical viewers. Whether super-recognisers also sampled more *useful* identity information compared to typical viewers is unclear. Therefore, in experiment 4, we aim to examine whether the *quality* of the identity information sampled from super recognisers differed from the information sampled from typical viewers.

Experiment 4 is a computational study designed to measure the quality of information sampled by participants during Experiment 3 and whether this alone can explain their superior face identity processing ability. Although Experiment 3 showed that super-recognisers and typical viewers sampled different face information, it is unclear whether differences in this sampled information are alone able to explain differences in their ability. To address this question in Experiment 4, we used Deep Convolutional Neural Networks (DCNNs) to quantify the identity information extracted by human observers in Experiment 3.

DCNNs provide stable computational metrics to measure similarities between face identities in images. Thus, despite showing cognitive differences compared to humans (see Chapters 2 and 3), DCNNs allow us to measure the amount of identity information contained in the sampled information by super-recognisers or typical viewers. The main question that arises is if the information sampled by super-recognisers contains more identity information than the information sampled by control participants. That is, will the information from super-recognisers lead to improve identification performance in DCNNs? If so, then super-recognisers may be better at face recognition simply because they extract more useful identity information via eye movements. If not, it would suggest that differences in super-recognisers' ability stem from the perceptual processing of the sampled information.

We set out to answer this question by combining gazemap datasets from Experiments 3A and 3B. Due to the constraints imposed by the COVID-19 pandemic, we opted to reuse data from Experiments 3a and 3b for this computational study. The pandemic's limitations on conducting new experiments made it more practical and efficient to leverage the existing data to address our research objectives. We use fixation coordinates from participants in this study to create a static representation of participants' perceptual sampling while observing faces. Our idea is to create a face-matching task where one of the pair images will be a static representation of the perceptual sampling made by a human participant on a given trial from Experiment 3 (from either a typical viewer or super-recogniser). The other image is a high-quality image showing the same person or different people. By measuring DCNN accuracy when matching these pairs, we provide a metric of the information value of identity information sampled from the face by the human participant.

In addition to comparing the information sampled by the two groups of participants in this way, we will also compare these groups to randomised fixation coordinates. Adding a randomised group is crucial because it will enable us to understand if human-guided identity information is more valuable than randomised information *in general*. We would expect this to be the case given that appears to be some correspondence between features used by humans and DCNNs for face identification (e.g. Abudarham et al., 2019). But proving this is the case is also important to inform whether the use of 'attention layers' in DCNNs that emphasise information that is useful for human participants (e.g. see Lai et al. 2020; Rong et

al. 2021; Yang et al. 2022). Thus, Experiment 4 will allow us to observe how much human attention can benefit DCNN accuracy compared to randomised information and if super-recognisers extract more valuable information than typical viewers.

Methods

Deep Convolutional Neural Networks (DCNNs)

We used nine different DCNNs in this study to measure the computational value of face identity information sampled by participants. These DCNNs were based on 5 different architectures trained on various datasets and implemented them using Keras (Chollet et al., 2015) or Pytorch (Paszke et al., 2019) in Python. Details of the training and architecture of these 9 DCNNs are shown in Table 4.4. All models were official models collected from the system developers (e.g. GitHub repository).

DCNN	Architecture	Dataset	Python Library
1	ResNet50 (He, Zhang, Ren & Sun,2016)	VggFace2 (Cao et al., 2018)	Keras
2	ResNet34 (He, Zhang, Ren & Sun, 2016) (https://github.com/ageitgey/face_recognition)	VggFace (Simonyan and Zisserman, 2014); Face Scrub dataset (Ng & Winkler, 2014) and images from the internet (King, 2009)	Keras
3	ResNet50	VggFace2	Pytorch
4	ResNet50	MS-Celeb-1M dataset (Guo et al., 2016) fine-tuned on VggFace2	Pytorch
5	Se-ResNet50 (Hu, Shen, & Sun, 2018)	VggFace2	Pytorch
6	Se-ResNet50	MS-Celeb-1M dataset fine-tuned on VggFace2	Pytorch
7	VGG16 (Simonyan and Zisserman, 2014)	VggFace	Keras
8	FaceNet (Schroff, Kalenichenko, & Philbin, 2015)	VggFace2	Pytorch
9	Facenet	CASIA-WebFace (Yi, Lei, Liao & Li, 2014)	Pytorch

Table 4.4 Description of architectures, datasets, and Python libraries used in this study. We also enumerated DCNNs from 1 to 9.

Stimuli

We used the fixation data from Experiment 3 to make images for the face-matching task in this study. In Experiment 3, we calculated the fixations of 68 typical viewers and 37 super-recognisers processing a face recognition task. During the study, participants

processed 144 identities while their eye movements were being tracked, and we manipulated the amount of information participants could sample from every fixation through variable spotlight aperture sizes (See Experiment 3). Here, we aggregated the aperture sizes with collected fixations on the face images to create a static representation of their perceptual sampling.

To reconstruct participants' perceptual sampling processing faces, we first reconstructed the perceptual information they sampled on each fixation. For that, we used each fixation coordinate to create an image layer, showing the given spotlight aperture for the coordinate position. We repeated this process for all fixations. We then merged all the single apertured fixation layers into a single image, visualising all fixations combined.. However, to enhance the realism of human perceptual sampling, we incorporated a retinal filter (Targino Da Costa & Do, 2014) into our image reconstruction process. By convolving the resulting image with this mathematical filtering, we aimed to simulate the visual processing that occurs in the human retina, making our perceptual sampling more accurate and representative of human vision. It is important to note that the retinal filtering will modulate the perceptual efficiency (i.e. acuity) inversely proportional to the distance of the fixation position. For that, the work of Targino and Do (2014) shows a mathematical 'foveation' model that aims to replicate human vision loss. Therefore, the retinal filtering proposed by Targino and Do (2014) in the fixation images will simulate their perceptual experience. We used the retinal filter parameters' distance to screen' set to 650mm and the loss parameter $\Delta = 25$ to mimic human perceptual loss. The final 'foveated image' is a reconstruction of the total participants' information sampling while processing the stimuli. We show this process and a resultant image of Experiment 4 in Figure 4.18.

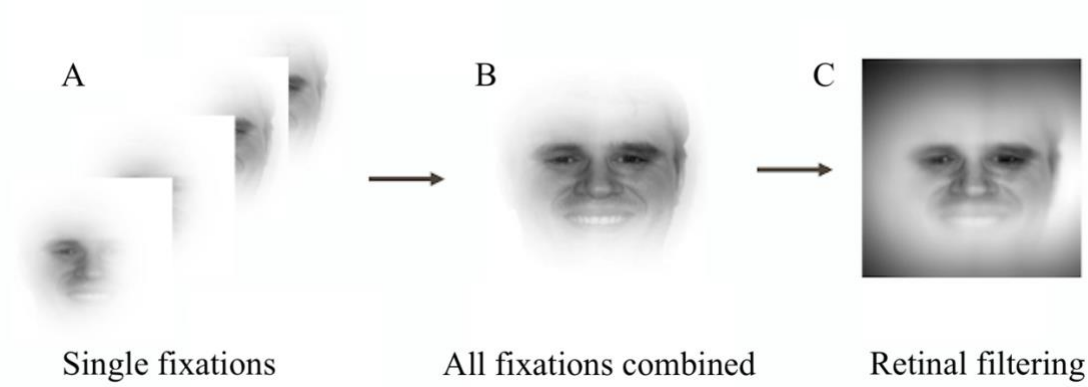


Figure 4.18. Illustration of how we created images used in Experiment 4. (A) Individual fixation coordinates were used to recreate a collection of image layers by convolving the viewed image region in the 'spotlight' aperture with a retinal filter centred at the point of fixation. (B) Image layers were then flattened to form a single image. (C) We applied a retinal filter (Targino Da Costa & Do, 2014) on the given fixations to exclude external information apart from the originally extracted by participants.

In addition to human-generated foveated images, we created equivalent foveated images for participants containing fixations in random coordinates. Each foveated image generated by participants had a 'yoked' random-generated image. This yoked image was generated in the same way as the human-generated images, with the same number of fixations, but using random fixation coordinates set within the bound of the face region. We used these images as a baseline for our analysis to investigate if human fixation patterns produce more computationally valuable face identity information than random samples of information. That is, if the foveated images produced by human fixations result in higher accuracy compared to the randomly generated ones, this will illustrate that the information sampled by humans is tuned to computationally relevant facial information for DCNNS. We show an example of the resulting foveated images per aperture size for both human and Randomly generated fixations in Figure 4.19.

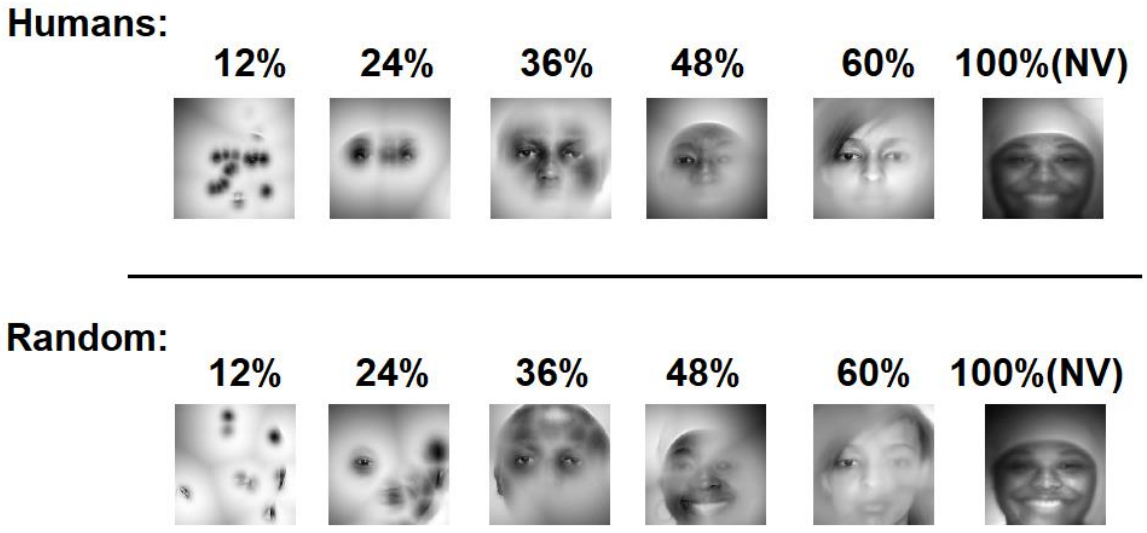


Figure 4.19. Representative examples of foveated image per aperture size generated by human-guided or randomly generated fixation positions.

Altogether, the process of recreating participants and randomly generated foveated images resulted in a total of 22,108 images divided between 3 groups: Super-recognisers,

Typical viewers, and Random. We used these images to create 44,216 pairs of images for the face-matching task, with 22,108 match and 22,108 non-match pairs. For the match pairs, one image was a foveated image and the other a different high-quality image of the same identity. For the non-match trial pairs, one image was a foveated image, and the other was a high-quality image of someone from similar demographics (e.g. same age, sex, ethnicity, etc.) but not the same identity. We manually selected pairs of similar identities for the non-match trials.

We used DCNN algorithms to generate similarity ratings for all the stimuli images. Thus, all stimuli images we used in Experiment 4 were 224x224 pixels containing an aligned face. To do this, we detected the facial regions of the stimuli images using a Multi-task Cascaded Convolutional Neural Network (MTCNN) (Zhang, Zhang, Li, & Qiao, 2016). The MTCNN detected, extracted, and aligned each face in the image set. We then resized all images to 224 × 224 pixels to serve as the DCNNs' input (see Cao et al., 2018) and converted them to grayscale.

Analysis

We examined the face-matching accuracy of DCNNs performing the stimuli described above. Similar to Chapters 2 and 3, we used the penultimate layer of DCNNs to generate a numerical description (i.e. a feature vector) for every image used in the stimuli. We calculated the Euclidean Distance between feature vectors to determine the similarities between image pairs. We used the inverse-normalised similarities as 'similarity scores' so that lower values signal different identity (i.e. non-match) pairs, and higher values signal the same identities (i.e. match). We report the accuracy of DCNNs as the Area Under the ROC Curve (AUC) computed from these similarity scores. We calculated a total of eighteen AUC scores for each DCNN algorithm used in this study. These AUC scores resulted from DCNNs analysing participant sources (typical viewers, super-recognisers or random) at each aperture size condition (12%, 24%, 36%, 48%, 60%, 100%).

To analyse these AUC scores, we used Linear Mixed Models (LMM) with a quadratic fit (aperture size^2) from the GAMLj module package in JAMOVI (version 2.2.5; The Jamovi project, 2021). We added aperture size (12%, 24%, 36%, 48%, 60%, 100%) and participant source (control, super-recogniser, and random) as predictors in the model. We used

Aperture size as a continuous variable and DCNNs as random effects. In addition, we ran a General Linear Model (GLM) with a quadratic fit (aperture size^2) to investigate whether the randomly generated foveated images revealed different amounts of information compared to the human-generated ones. We used a quadratic fit in these analyses to avoid ceiling effects derived from the natural viewing condition.

Results

Overall Accuracy

We show the accuracy of DCNNs processing the stimuli in Figure 4.20. In Figure 4.20, we show the calculated AUC for all 9 DCNNs processing stimuli images separately for each of the six aperture conditions and colour coded the three participant sources. Unsurprisingly, inspection of Figure 4.20 shows that the constrained information caused by smaller apertures causes reduced accuracy of DCNNs for all participant sources. To analyse this effect, we used linear mixed model analysis. For the model of Accuracy (AUC), we set DCNNs' intercept as a random effect and aperture size, aperture size^2 (quadratic fit), and source (human or random) as fixed effects. The linear mixed model reveals a significant main effect of aperture size [F(1,147)= 770.91, b = 0.015, CI = [0.014, 0.016], t(149) = 27.23, p < .001] and aperture size^2 [F(1,147)=388.67, b = -9.15e-5, CI = [-1.01e-4, -8.23e-5], t(149) = -19.32, p < .001], meaning that for larger aperture sizes DCNNs were better able to extract valuable identity information.

More interestingly, we found a significant main effect of source (Typical Viewer, Super-Recogniser, or Random) [F(2,147)=21.77, p< .001], with fixed effects showing a significant difference between Random and Typical viewers [b = -0.05, CI = [-0.076, -0.028], t(147) = -4.26, p< .001] and Super-Recognisers and Typical Viewers [b = 0.027, CI = [0.003, 0.052], t(147) = 2.23, p = 0.027]. This significant difference is interesting because it illustrates that DCNNs benefit from human-guided attention to faces, and super-recognisers' information is the most useful. We found no significant two-way interaction in this analysis. See APPENDIX C – Experiment 4 for the full model table.

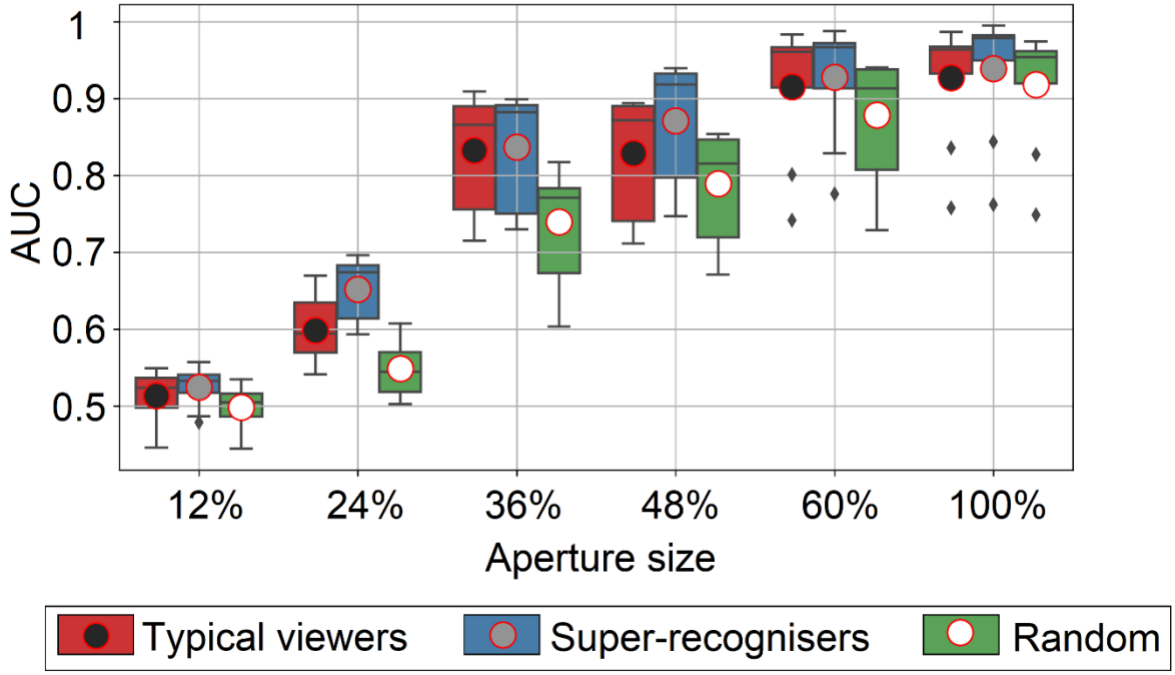


Figure 4.20. Accuracy (AUC) of DCNNs performing the stimuli in the six aperture conditions for the three participant groups.

Information available

In the previous analysis, we found that super-recogniser-guided information provided more useful identity information for DCNNs. However, Experiment 3 found that super-recognisers showed enhanced exploration of faces compared to typical viewers. Thus, the enhanced exploration of super-recognisers might have caused the advantage of DCNNs in using their information. We then calculated the total information available for each image used in this study. For that, we replicated the foveation procedure (see Methods - Stimuli) but used a blank white image instead of a face image. After foveation, we calculated the proportion of revealed pixels as at least 80% white. We used this proportion (in percentage) as the amount of information available for each foveated image.

We show the resulting information available for DCNNs processing the stimuli in Figure 4.21. Similar to the previous figure, we divided the resulting information separately for the six aperture conditions and colour-coded the three participant sources. Visual inspection of Figure 4.21 reveals that smaller aperture sizes -on average- limited the amount of information available, and Random-guided fixations produced more available information than human-guided ones. We investigated these assumptions using Generalised Mixed

Models. For the model of available information, we investigated the effect of participant source (Typical Viewer, Super-Recogniser, and Random) and Aperture size[2] as factors of the model. Not surprisingly, generalised mixed models showed a significant main effect of using quadratic fit (aperture size[2]) [F(1,44209)= 51433, η^2p = 0.538, *b* = -0.016, CI = [-0.017, -0.016], *t*(44209) = -226.8, *p* < .001], showing higher information available for bigger apertures. In addition, this analysis revealed a significant main effect of Group (Super-Recognisers, Typical Viewers, and Random)[F(2,44209)= 1840, η^2p=0.077, p< .001], with fixed effects showing higher information available for Random-guided fixations compared to typical viewers [*b* =10.96, CI = [10.58, 11.35], *t*(44209) = 55.45, *p* < .001]. In addition, fixed effects also show that Super-Recognisers explored facial regions more than Typical Viewers [*b*= 3.29, CI= [2.856, 3.724], *t*(44209)= 14.87, *p*< .001]. The fixation patterns of humans revealed less visual information than that generated by random fixation positions, showing that the accuracy advantage for human-generated images was due to quality rather than quantity of face information. Still, the enhanced exploration of Super-Recognisers demonstrates that they investigate more areas possibly containing relevant identity information than typical viewers. See APPENDIX C – Experiment 4 for the full model tables.

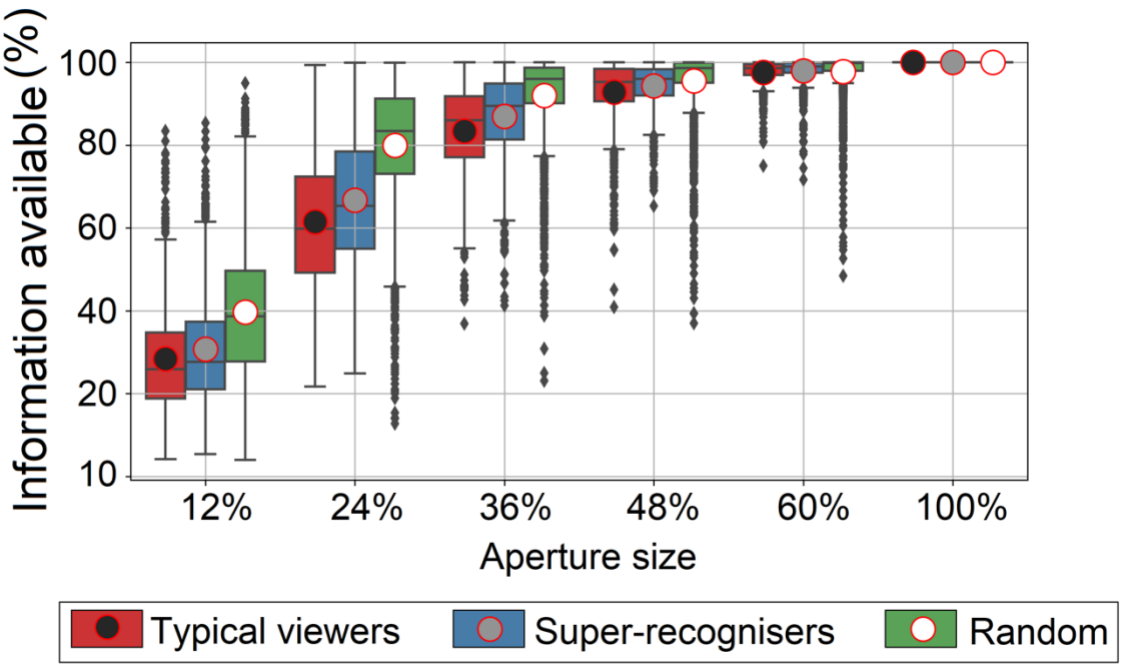


Figure 4.21. Image information available across aperture conditions for the three participant groups.

Discussion

Experiment 4 was a computational study investigating the quality of face information extracted by super-recognisers and typical viewers compared to a baseline of random samples. We performed this experiment because, in the previous experiment (i.e. Experiment 3), we found significant differences between the fixation patterns of super-recognisers and typical viewers investigating faces. Therefore, it was still unclear how much these fixation pattern differences would reflect in the quality of identity information extracted by participants. So we reconstructed the perceptual sampling experience of human participants (and a random baseline) using their fixation coordinates and retinal filters and used these images as input to DCNNs in an identity-matching task.

We used 9 DCNNs to investigate the quality of identity information extracted by humans and randomly placed fixations in the foveated images. If DCNNs could correctly identify someone's identity using the foveated stimuli images, this shows that the samples contained valuable identity information. This enables us to obtain an objective measure of the amount of computationally useful identity information contained in the perceptual sampling of our participants in Experiment 3.

The first result was that the reconstructed images generated by human perceptual sampling provided better identity information than random samples of face information. This was despite random samples revealing more information. This shows that despite some differences between humans and DCNNs processing faces, as shown in previous chapters, there is an overlap in the information that humans and DCNNs use to identify faces (see Lai et al. 2020; Rong et al. 2021; Yang et al. 2022). The second result was that reconstructed images generated by super-recognisers samples produced significantly higher accuracy than typical viewers'. Therefore, we can conclude that super-recognisers extracted more valuable identity information compared to typical viewers.

Our results show that the method developed for Experiment 4 revealed to be an interesting approach to quantifying identity information in diverse 'foveated images'. In Experiment 4, we used the perceptual sampling provided by humans or randomised fixations and found that DCNNs improve their accuracy when humans produce such sampling. Thus, as aforementioned, such results would indicate –at least- commonalities in how DCNNs and humans process identity verification. However, future studies should

address the question regarding their similarities starting from the other direction. That is, because the method described here is a computational evaluation, researchers could first find the set of – for example – one or two optimal features that DCNNs most benefit from for identity verification and later compare if humans also agree with the decision. Such an approach would give us strong evidence to assess how DCNNs and humans engage similarly in processing faces and further respond if they share similar mechanisms in doing so. More, this presents a compelling opportunity for future research to delve into how these facial region similarities between DCNNs and human could be leveraged to enhance the fusion between their decisions (e.g. Chapter 3). By exploring how the convergence in decision-making processes can be harnessed to improve the fusion model, researchers can unlock new possibilities for creating more accurate and efficient forensic systems that better use the strengths of both humans and computational algorithms processing identities. Furthermore, it is crucial to highlight that our research objective focused specifically on investigating facial areas, which led us to exclude entirely any fixations that landed outside of the facial region. This deliberate exclusion opens up an intriguing opportunity for future studies to explore the interaction between these disregarded fixations and facial recognition performance. Understanding how fixations outside the facial region may impact facial recognition processes could provide valuable insights into the broader context of visual perception and its implications for identity verification tasks.

Chapter discussion

In this chapter, we used eye-tracking devices to investigate how humans and DCNNs use featural information from faces while performing face-matching and recognition tasks. We divided this chapter into three main experiments (Experiments 2, 3, and 4). In Experiment 2, we investigated the use of featural information differences between human participants performing a face-matching task. In Experiment 3, the use of featural information differences between human participants performing a face recognition task. And in Experiment 4, we investigated if the information sampled from participants -during Experiment 3- were beneficial for DCNNs. We used this body of experiments to clarify how individual differences relate to the perception of identity information in faces.

In Experiment 2, we investigated participants' fixation patterns leading to improved face-matching performance while processing a face-matching task. In this experiment, we introduced the PCA+MLDA method to visualize the differences in fixation patterns between two groups of participants with distinct face-matching abilities (high and average ability). We also compared the results with conventional ways to investigate differences in fixation patterns. Using the conventional way to analyse fixation patterns, we found that greater attention to the right eyes (of both faces of the matching pair) predicted face-matching ability. Our proposed PCA and MLDA methods also highlighted the right eyes as a predictor of performance. However, the PCA and MLDA helped us enrich our understanding compared to conventional methods. This new method helped us visualise that face exploration and an overall tendency toward the right side of faces (i.e. not just eyes) predicted performance in face-matching tasks.

In Experiment 3A, we investigated participants' fixation patterns leading to improved face recognition performance. We performed this investigation using the PCA method introduced in Experiment 2. The PCA significantly explained distinct fixation patterns between super-recognisers and typical viewers performing the task. Our visualisation method shows that super-recognisers explore more facial regions while avoiding the eye region. Typical viewers, on the other hand, showed less face exploration and more attention to the eye region. Interestingly, we found that these differences in sampling strategy are more prominent when participants learn the identity depicted in faces. This higher exploration found in super-recognisers raises the argument that they show more efficiency in absorbing face information as they had only 5 seconds to learn the identity of faces.

The information sampled by super-recognisers when learning or recognising faces is not qualitatively different from typical viewers' because we replicated the PCA model found in Experiment 3A using a cohort of majorly typical viewers in Experiment 3B. In Experiment 3A, we used two well-established groups of typical viewers and super-recognisers. As aforementioned, such investigation using the PCA approach revealed substantial differences between the two groups in the fixation patterns when learning and recognising faces. These differences were apparent in the first principal components. However, it was still unclear if the PCA model reflected qualitative differences between the two groups or simply quantitative differences in the information sampling strategy that would be found in any cohort of humans. And so, our objective with Experiment 3B was to investigate the stability

of the calculated PCA model when using a cohort of majorly typical viewers. Our results show striking similarities between the PCA models from Experiments 3A and 3B. Thus, such robustness in the PCA model clarifies that the different strategies engaged by super-recognisers when learning or recognising faces are not qualitatively different from the ones engaged by typical viewers.

The results found in Experiments 2 and 3A/3B are not contradictory. Our analysis for both experiments suggested that the facial information sampled reflects individual differences in face processing ability. In Experiment 2 (i.e. a face-matching task), we found that those who attended more to the eye region possessed a higher ability. In contrast, Experiment 3 (i.e. a face recognition task) shows that those who avoided the eyes possessed a higher ability. It is important to address the fundamental differences in both experiments, as Experiment 2 required no memory component. Thus, the results found between them are not necessarily contradictory. First, it outlines that high-performers possess high levels of adaptability depending on the task they process. Second, in the two distinct experiments, we found that high-performers sample more facial information from the stimuli. And so, our investigation to verify the links between holistic processing and individual differences (e.g. Bobak et al., 2017) was not conclusive. Instead, our analysis revealed that high performance in face identity processing relates to the higher exploration of facial features and accurate changes in information sampling for adequate processing.

In Experiment 4, we asked whether the information sampled from super-recognisers would contain more quality of identity information than information sampled from typical viewers and randomised information. For this analysis, we collected participants' fixation patterns from Experiments 3A and 3B and simulated their perceptual sampling experience as a static foveated image. We also created a version of such foveated images using randomised fixation positions. In Experiment 4, we used DCNNs to measure the quality of identity information. We used DCNNs to generate similarity scores for foveated images paired against other images of the same person or different people. When using such similarity scores, a higher accuracy would directly reflect a higher quality of identity information. Ultimately, our results show that identity information required some level of organisation as randomised information provided lower accuracy for DCNNs. Interestingly, and similar to previous studies (e.g. Abudarham, Shkiller, & Yovel, 2019), DCNNs seem to use information similarly to humans coding for identity due to their superior precision when

using their sampling. Interestingly, super-recognisers provided the most useful identity information. This result extends the previous Experiments 2 and 3 because it shows that super-recognisers not only explore more facial areas but also attend to diverse areas containing valuable identity information for further recognition.

It is still unclear if the information sampled by super-recognisers would reflect better face identification performance in – for example – typical viewers. In Experiment 4, we investigated the quality of identity information by using DCNNs and found that the perceptual sampling from super-recognisers resulted in better face identification performance. However, Experiment 4 was a computational study, so it is still unclear if humans would agree with such results using DCNNs. For instance, future studies should create pairwise matching or recognition tasks where human participants explore face identity information through the perceptual sampling of super-recognisers, typical viewers, and randomised fixations. Notably, such tasks would be created by using foveated images using the methods described in Experiment 4. Such investigation would reveal if humans can improve - or decrease - their face identity processing performance by different perceptual sampling information.

Altogether, the three experiments in this chapter helped us visualise that super-recognition might relate to extracting quantitatively more valuable information from faces. Our results demonstrate a more local/featural processing in enhanced face processing ability. Although past studies show that super-recognisers require less local information to achieve accurate recognition (Royer et al., 2015; Tsantani et al., 2020) and their superior ability often relates to enhanced holistic processing (e.g. Bobak et al., 2017; Bennetts, Mole & Bate, 2017). Here, we provide evidence that super-recognisers show efficiency and flexibility using diverse valuable local identity information from faces. Thus, our results suggest that they use this compendium of sampled local information to create a better representation of faces, opposite of previous work suggesting that individual differences in identity processing might be related to enhanced holistic processing. However, it is important to note that the stimuli processed by human participants throughout Chapter 3 were cropped photographs of faces. Therefore, it remains unclear if such mechanisms for processing identity verification would reflect the same mechanisms found -for example- when individuals have an 'in the wild' face-to-face interaction with other individuals. That is, previous work suggests that the mechanisms when processing screen-based stimuli are

significantly different compared to when processing the same stimuli outside the computer (e.g. Nasiopoulos, Risko & Kingstone, 2015). And so, in Chapter 4 and 5 of this thesis, we will assess if the mechanisms of attention -measured by eye fixations and mobile eye trackers- would be similar compared to what we found here in Chsssssapter 3.

Chapter 5 - Using DCNNs to examine human face perception in the wild

Throughout my thesis, I have applied computational methods borrowed from the field of computer vision to examine individual human and DCNN performance on face identification tasks. In the previous chapters of this thesis, we investigated humans processing facial images on computer screens in the lab. In chapter 5, I develop a new innovative technique for the analysis of wearable eye-tracking technology, which uses automatic person-detection software to measure social attention. This enables us to investigate human processing faces in outdoor and uncontrolled environments. This chapter is adapted from a full research paper, which has been accepted for publication in Nature Scientific Reports, and a pre-print is available online (Varela, Towler, Kemp, & White, 2022).

Introduction

One approach to studying person perception is to examine socially-directed attention by analysing people's eye movements as they view images of people presented on computer screens (e.g. Yarbus, 1967; Amso, Haas & Markant, 2014; Birminghan, Bischof & Kingstone, 2009; Bobak et al., 2017; Rösler, End & Gamer, 2017; Gregory, Bolderston & Antolin, 2019). However, photographs of social scenes do not represent the dynamic, multidimensional reality of our social experience. Indeed, participants fixate on faces less in face-to-face interactions than when watching video stimuli (Nasiopoulos, Risko & Kingstone, 2015; Risko, Richardson & Kingstone, 2016; Laidlaw et al., 2011; Foulsham, Walker & Kingstone, 2011), indicating that contrived laboratory tasks are inadequate analogues of real-world social attention (Kingstone, 2009; see also Nasiopoulos, Risko & Kingstone, 2015; Risko, Richardson & Kingstone, 2016).

Surprisingly little is known about how people direct their attention towards others in naturalistic social environments. Yet, this information provides valuable constraints to understanding the perceptual processes and mechanisms of attention. For example, researchers have captured the visual experience of babies and toddlers using wearable cameras, enabling researchers to better understand how perceptual expertise with faces develops. This work shows that faces are present in infants' field of view roughly 25% of the time (e.g. Sugden & Moulson, 2019), with the vast majority of this exposure to familiar faces of primary caregivers. In contrast, faces make up a far smaller fraction of children's visual

experience beyond their first birthday (~10%, e.g. Jayaraman et al. 2015; Fausey et al. 2016). The extent to which babies and children attend to these faces is less clear, but quantifying the frequency of faces in the field of vision, and the categories of faces making up this exposure, informs theories of perceptual expertise by grounding them in the information available in the visual environment (e.g. see Young & Burton, 2018).

Studies of adults' attention to people in natural settings are extremely rare, and almost all knowledge on this topic comes from tightly controlled laboratory-based research. This laboratory-based research shows, for example, that faces capture attention and are processed preferentially relative to non-face objects and bodies (Bindemann et al., 2005; Theeuwes & Van der Stigchel, 2006; Morrisey et al., 2019), and this leads to the view that this process operates automatically (see Palermo & Rhodes, 2007 for a review; Yan, Young & Andrews, 2017). However, it is not clear whether this holds for ambient environments populated with many competing stimuli – each with its unique affordance (Gibson, 1979) – and where the 'social stimuli' are real people, complete with legs, minds, and eyes of their own.

These knowledge gaps have increased interest in methods that allow studies of person perception and social attention in complex environments. One approach has been to use virtual reality, with faces rendered on animated bodies in virtual worlds (Bindemann et al., 2021; Fysh et al., 2021; Bülthoff et al., 2019). Another has been to study social attention in 'the wild' by studying the eye movements of participants wearing eye-tracking devices that monitor their fixations as they navigate real-world ambient environments (see Foulsham, Walker & Kingstone, 2011; Foulsham, 2020; Tatler, Hanzen & Pelz, 2019 for recent reviews). Wearable eye-tracking offers the advantage of studying social attention and person perception *in situ*. However, it requires experimenters to view long video recordings and manually code what is being fixated in every frame of video recorded. Even coding simple aspects of gaze fixations, for example, counting person fixations vs non-person fixations, is extremely time-consuming (Foulsham, Walker & Kingstone, 2011; Hessels et al., 2020; De Lillo et al., 2021), making the examination of social attention in naturalistic environments impractical at the resolution afforded by lab-based eye-tracking studies (e.g. Bobak et al. 2019; Rice et al. 2013; Yarbus, 1967), and experimenters are limited to coarse analysis of fixation patterns.

Experiment 5

In Experiment 5, we introduce a novel method that enables fine-grained investigations of naturalistic social attention for the first time. Our 'dynamic regions of interest' (dROI) approach automatically measures social attention in ambient environments frame-by-frame. We achieve this by co-registering eye-movement data from a wearable eye-tracker with body and face landmark positions extracted from video data using a state-of-the-art computer vision algorithm (Cao et al., 2019). This method automatically encodes eye fixations directed towards people before mapping the locations of these fixations to landmarks on the face and body. Our approach overcomes many significant limitations of prior work on social attention in naturalistic environments, saving substantial research effort by avoiding the need for manual coding of fixations to pre-specified regions (e.g. Mele & Federici, 2012; Benjamins, Hessels & Hooge, 2018; Hessels et al. 2020; Haensel et al., 2020; De Lillo et al., 2021). In addition to removing the burden of manual coding, our approach also increases temporal resolution and the volume of data, enabling new analytic approaches which open up new avenues to study person perception in unconstrained environments.

Given this is the first piece of work to use this approach, we have addressed some preliminary research questions to demonstrate its diverse applications. First, we quantify the extent to which people attend to the bodies and faces of passersby and ask whether faces do actually 'capture' viewers' attention as previously claimed. Second, we conducted an exploratory analysis to visualise the appearance of faces that participants chose to fixate on compared to those they did not.

Third, we ask whether patterns of social attention 'in the wild' reflect stable individual differences in observers, both when participants were walking in a public space and when they were engaged in face-to-face social interaction. Recent eye-tracking studies have shown large individual differences in the way that people attend to social scenes shown on screens (Constantino et al., 2017; Kennedy et al., 2017), and abnormal social attention is associated with broader social deficits, for example, in Autism Spectrum Disorder (see Guillon et al. 2014). These studies also point to a genetic basis underlying people's social attention (Constantino et al., 2017; Kennedy et al., 2017), but no study has measured individual differences in complex everyday environments. Similarly, studies have

found significant and stable individual differences in people's face-processing abilities (White & Burton, under review), and these are associated with different patterns of eye movements to faces and people in lab-based tasks (Bobak et al. 2017; Dunn et al., 2022), but it is not clear whether these differences transfer outside of the lab to more naturalistic perceptual environments (see Ramon, Bobak & White, 2019).

Methods

Participants

Thirty-three university students from UNSW Sydney completed the study in return for course credit (9 males, 24 females; Age M= 21.4, SD= 5.4). We excluded full data from two participants because of corrupt eye-tracking data. In addition, procedural issues meant that we deleted data from one segment of the navigation task (see below) for one participant, and we deleted data for the face-to-face task (see below) for three participants. This gave a total of 31 participants in the main navigation task analysis, 30 in the individual differences analysis of the navigation task and 28 in the face-to-face interaction analysis.

Apparatus and eye movement classification

We used a wearable eye-tracking device to record participants' eye gaze data as they completed the study (Pupil Labs Core: Kassner, Patera & Bulling, 2014). This device recorded videos of participants' field of view and eye gaze coordinates. A set of three cameras achieve this recording, a frontal camera facing the environment and two cameras facing the eyes. The resolution of the frontal camera was 1920x1080 pixels at 60 frames per second, and the cameras facing the eyes were both of resolution 192x192 pixels at 120 frames per second. The wearable eye-tracker was connected via USB to a laptop (Dell XPS 13 7390 2-in-1 placed inside a backpack worn by the participant. We used Pupil Capture to save video and eye gaze data, and Pupil Player to calculate fixations (Kassner, Patera & Bulling, 2014).

Procedure

We conducted the study during term time when the campus was busy. There were no COVID-19 cases in Sydney at the time and so people were not wearing facemasks. We fitted

the mobile eye-tracking device to participants (see Materials above). Participants then completed two tasks while the device recorded their eye gaze: a face-to-face interaction task, where they interacted with the experimenter for a brief period; and a navigation task, where they walked around the UNSW Sydney campus following a circular route.

In the face-to-face interaction task, participants stood in an empty corridor, directly facing the experimenter at a distance of 1.5m (see Figure 5.5A). Participants listened to verbal instructions provided by the experimenter about the navigation task, explaining that this was a naturalistic study and that they should walk through campus as they would on a normal day. The experimenter explained the route they should take before asking participants if they understood and had any questions before beginning the task. The experimenter delivered these instructions by reciting a pre-defined script, and participants spent an average of 30 seconds listening and asking questions about the task. This recording was used in the analysis 'Fixation patterns during face-to-face interaction associated with face recognition ability'.

Participants then followed the experimenter to a separate room where a map and pictures of the walk were on the wall showing the study route. When participants indicated they were ready to begin, participants exited the room with the experimenter, and the Navigation task began. Participants navigated a pre-defined circular route via the main campus thoroughfares passing busy places (e.g. coffee shops, library, food court) through indoor and outdoor settings. Participants were always under the experimenter's supervision, who kept a ~2.5m distance behind participants. When participants arrived at the library, we asked them to stop walking and rest for a minute, which divided the study route into two segments. Segment 1 lasted approximately 12 minutes on average, and segment 2 approximately 4 minutes.

After completing the wearable eye-tracking tasks, participants completed a standard measure of unfamiliar face memory ability, the Cambridge Face Memory Test extended version (CMFT+) (Russell, Duchaine & Nakayama, 2009) and a self-report measure of face recognition ability, the Prosopagnosia Index short version (PI-20: Shah et al., 2015). This CFMT+ asks for participants to learn and memorise the grayscale faces of 6 caucasian males to be recognised later in 102 three-alternative trials without any time limit. The CFMT+ is a challenging test because the learned faces change in angle of view and image quality in the trials. The PI-20 is composed of 20 questions such as "My face recognition ability is worse

than most people", and participants must rank their responses from "Strongly agree" to "Strongly disagree". Participants were also asked questions during the debriefing to gauge their awareness of the study's purpose. Only four participants mentioned attention to people or person perception as a potential research topic.

Eye gaze data processing

The eye-tracking device collected raw gaze data of participants. We transformed this data into fixations, saccades and blinks using open-source tools provided by the eye-tracking manufacturer (Pupil Capture and Pupil Player, see https://pupil-labs.com/products/core/). In all cases, we used the default settings. Fixations were output as coordinates labelled to specific pixels on the frontal camera frames. For analysis, we only considered frames with fixations.

Our main methodological advance was to automatically detect the presence of people in the participant's field of view using open-source body and face detection tools (OpenPose: Cao et al., 2019). This tool detects people in video frames and automatically estimates up to 25 landmarks on the body and 70 on the face (if the person is sufficiently close to the viewer). Co-registering fixations with these landmarks enabled us to construct detailed maps of participants' attention to people. Interestingly, by using our approach, we could analyse a vast amount of fixation data without compromising the privacy or confidentiality of any given fixated person. This not only facilitated a comprehensive examination of the participants' visual behaviour but also ensured ethical considerations were upheld throughout our research. The individuals encountered during the experiment paradigm did not provide permission for their image use. However, their images were not used in paper figures, and their images were only used for a data processing pipeline that extracted locations of facial/body features but did not retain identifying information.

We used two methods to measure participants' attention to faces and people. In the first method (see Figures 5.1A and 5.4A), we registered fixations to the closest detected body or face landmark, considering only landmarks that OpenPose detected with a greater than 60% confidence. We chose a 60% confidence rate because our testing suggested this effectively excluded false positive 'phantom' bodies, which sometimes briefly appeared in the scene. We calculated the distance between fixation coordinates and landmark

coordinates for every frame containing both fixations and landmarks. Where the Euclidean distance between a fixation coordinate and the closest landmark was below a designated threshold, we registered a to that landmark. Thresholds varied depending on the spatial resolution of the landmark data being used (navigation task = 70 pixels; face-to-face interaction = 30 pixels). For the purpose of analysis in the navigation task, we clustered 25 landmarks into two categories (face and body; see Figure 5.1A), and for the face-to-face interaction task, we clustered 70 facial landmarks into five categories (nose, left/right eye, mouth, and the exterior of the face; see Figure 5.5A).

In the second method, we aimed to determine the precise location of fixations in a face to facilitate heatmap analysis in the face-to-face interaction (see top panel of Figure 5.6A). We achieved this by computing the relative position of a given fixation coordinate amongst facial landmarks using Delaunay triangulations (Delaunay, 1934) followed by Affine transformations. This way, fixation coordinates that landed within a given computed landmark triangle can be projected on the relative triangle in the standard template. This method enabled us to aggregate fixation data to more precise locations on the face to create a heatmap for each participant during the task.

Comparing automatic versus manual coding

We compared estimates of the number of people present in a video frame made by OpenPose (Cao et al., 2018) and humans. Four lab volunteers manually counted the number of people in each of the 560 randomly selected video frames from the navigation task (see Figure D.6 in Supplementary Materials for an example of a video frame). We found a very strong positive correlation between manual and automatic people counting ($r = 0.89$, $p < 0.001$, $n = 560$; see Figure D.7 in Supplementary Materials for scatterplot).

Navigation task

In the navigation task, we registered fixations as being to the head, body, or 'not-person' fixations. Head and body fixations were registered when OpenPose had greater than 60% confidence in either head or body regions and when a fixation was detected within 70 pixels of a landmark. Not-person fixations were any fixations that did not meet these

criteria. Probabilities of fixations to each of these three dynamic regions of interest (dROI) were calculated only for frames where a fixation was recorded. These data were filtered based on OpenPose detection as described in the Results section.

Face diet

To investigate the face diet of participants, we generated two average faces for fixated and not fixated faces (see Figure 5.4). First, we collected images of all faces OpenPose detected in each participant's video recording and stratified these according to whether the participant had fixated on them. We then used a face recognition algorithm (ResNet50 (He, Zhang, Ren & Sun,2016) trained on the VGGFace2 Database (Cao et al., 2018)) to find all instances of these fixated faces in participants' recordings. We achieved that by estimating the number of identities using K-means clustering and the Elbow method to find the optimal number of identities amongst fixated and not fixated faces. We then averaged all the images of each person's face to create an average per face identity and then averaged fixated and non-fixated faces separately to create the images shown in Figure 5.4.

Face-to-Face interaction task

For the face-to-face interaction task, we processed gaze data using landmark and heatmap registration methods. Participant heatmaps were analysed using principal components analysis (PCA) to identify major components (PCs) in the inter-individual variation of heatmaps, returning a set of PCs ranked according to their explained variance (see Chapter 4; Varela et al., 2018). The raw input data for the PCA is shown in Supplementary Materials (Figure D.4).

Results

Faces of passersby do not capture attention in live natural settings

Thirty-three participants followed a circular route around a busy university campus wearing a mobile eye-tracking device. We show an example video frame illustrating the eye-

tracking data provided by the eye-tracker and our detected dynamic regions of interest in Figure 5.1A (left panel). Our dynamic region of interest (dROI) analysis of social attention relied on automatic face and body detection algorithms developed by Cao and colleagues (OpenPose: Cao et al., 2019). We verified the accuracy of this algorithm on our video data by comparing its detections to manual coding of body presence by four human observers and found a high level of agreement (see Jongerius et al., 2021).

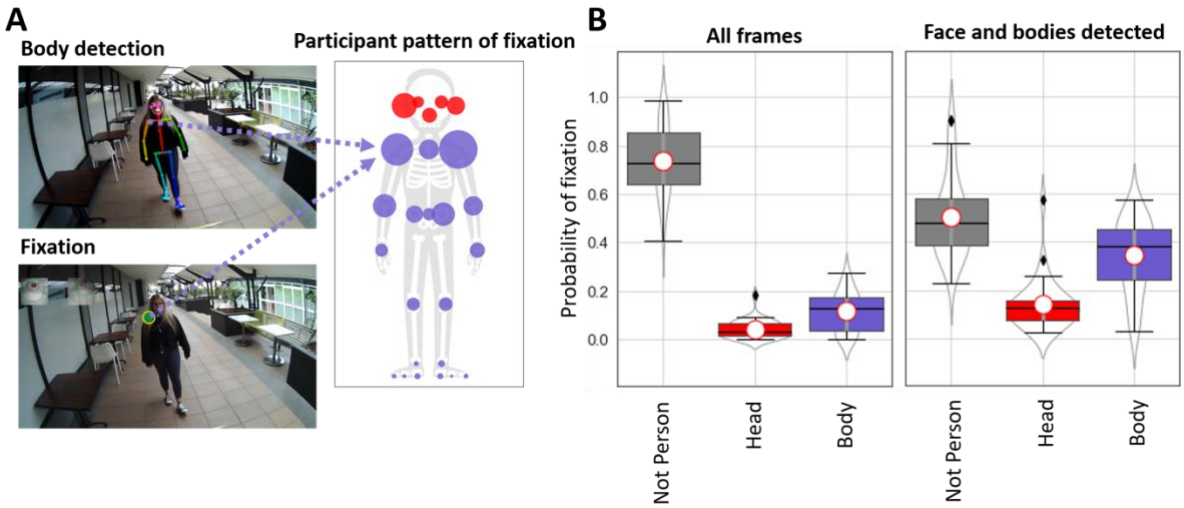


Figure 5.1. Dynamic region of interest (dROI) analysis of social attention while navigating a university campus. (A) Using data from a wearable eye-tracker, we extracted body landmarks from videos using OpenPose (top left) and co-registered viewers' fixations towards these landmarks (bottom left). The skeleton figure shows a participant's relative proportion of fixations to each body landmark, indicated by the size of the marker (all individual participant maps are available in Supplementary Material, Figure D.1). (B) The left panel shows overall proportions of non-person, head and body fixations as a proportion of all fixations in the recordings. The right panel shows boxplots of proportions of non-person, head and body fixations only as a proportion of frames where the algorithm detected heads and bodies. See main text for analysis.

To calculate the proportion of fixations participants made to faces and bodies, we co-registered fixation locations from the eye-tracker with landmarks on faces and bodies (Figure 5.1A, left; see Methods - Eye gaze processing). As a function of total fixations (Figure 5.1B, left), 16% of fixations were directed to people, with just 4% directed at people's heads (Body: M = 11.6%, SD = 8.3%; Head: M = 4.3%, SD = 3.8%). Restricting the analysis to only frames where faces and bodies were detected by the algorithm, we observed higher proportions of fixation towards people (50%), but fixations to heads remained relatively low at 14% of fixations (Body: M= 34.4%, SD= 14.9%; Head: M=14.4%, SD= 10.3%). The small proportion of fixations to faces may suggest that the widely reported finding that 'faces

capture attention' in lab-based studies (e.g. Vuilleumier, 2000; Bindemann et al., 2005; Theeuwes & Van der Stigchel, 2006; Gamer & Büchel, 2009; Ro, Russell & Lavie, 2001; Bobak et al., 2016; Rösler, End & Gamer, 2017; Gregory, Bolderston & Antolin, 2019) is not reflective of what occurs when we encounter unfamiliar people in public spaces.

Previous lab-based studies have shown that frontal faces capture more attention than averted faces (e.g. Shirama, 2012; Palanica & Itier, 2015). In a final test of whether unfamiliar faces capture attention in naturalistic settings, we compared the proportions of fixations to people in frames where the algorithm detected full faces (i.e. all facial features) against when the algorithm detected partial faces (i.e. a subset of facial features). This provided a test of whether frontal faces are fixated more than averted faces in a live natural setting.

Figure 5.2 shows that participants made more fixations to heads when full faces were visible compared to when faces were partially visible (Full = 15.1%, Partial = 12.3%, $t(30) = 2.58$, $p = 0.015$), but they also made more fixations to bodies when faces were fully visible (Full = 36.6%, Partial = 32.1%, $t(30) = 2.58$, $p = 0.015$; see Supplementary material for full ANOVA). This result suggests that participants were more likely to fixate on *people* when their faces were in full view but provides no evidence that faces captured this attention any more than other body regions.

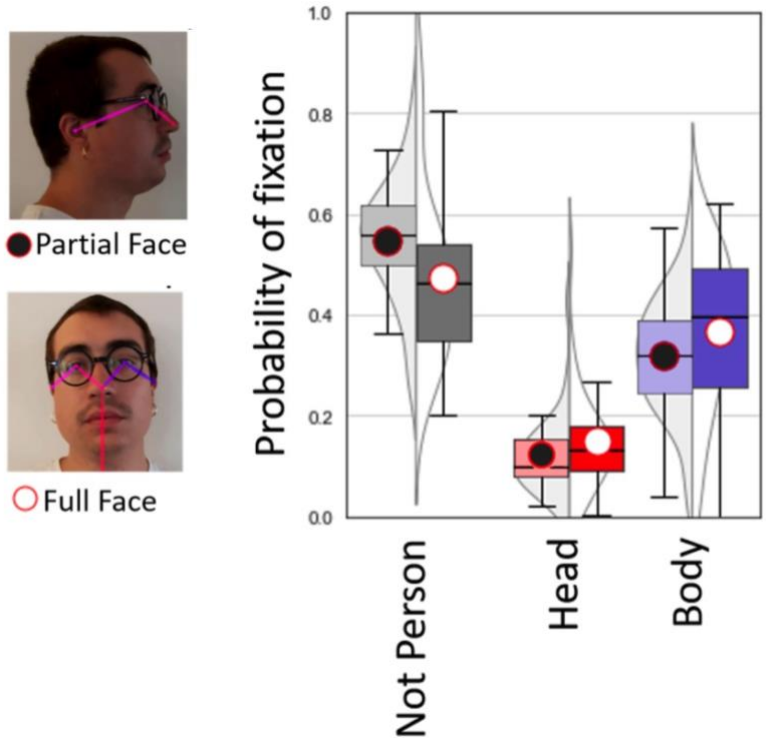


Figure 5.2. Comparing social attention when faces are fully visible versus partially visible in a video frame. Partially visible faces were due to non-frontal head angle or occlusion (figure legend, top). Results show a greater probability of fixating on people (heads and bodies) when their faces were fully visible. See the main text for analysis and Supplementary Materials for full ANOVA.

Overall, while in contrived lab-based tasks faces are attended to more (e.g. Bobak et al. 2017) and capture more attention (e.g. Theeuwes & Van der Stigchel, 2006) compared to other visual objects, using our technique to study social attention in a natural environment we did not find evidence that faces of passersby receive prioritised processing.

Influence of social attention on 'face diet'

Face 'diet' refers to the volume and composition of a person's perceptual exposure to faces. This concept has theoretical influence because exposure to faces is argued to underpin people's specialised expertise in processing faces (e.g. see Rhodes et al. 2005). Face diet tends to be made up of faces that are from similar demographic groups to our own, and so this concept has been used to explain the 'other-race effect' whereby people are better at recognising own-race faces than other-race faces (e.g. Crookes & McKone, 2009; McKone et al., 2019). But face diet is typically conceived as passively 'absorbed', and

the influence of a person's social attention on their face diet in natural settings is still unclear.

In Figure 5.3, we show how combining automatic face detection with wearable eye-tracking can be used to measure differences in participants' face diets. Using our dROI approach, we extracted images of faces and feature locations from video frames – both for fixated and non-fixated faces. This approach enabled us to generate image averages of fixated and non-fixated faces via an image morphing procedure (see Methods – Data analysis). The average appearance of fixated and non-fixated faces in Figure 5.4A shows subtle differences in expression and skin tone, suggesting that face diet is – perhaps – shaped by social attention in systematic ways. Moreover, Figure 5.4B shows that most fixations to faces concentrate at the centre of the participants' field of view. This result could mean that participants turned their heads to look at faces directly. Alternatively, participants may be more likely to look at people's faces when they pass directly in front of them (see Solmon, Foulsham & Kingstone, 2017).

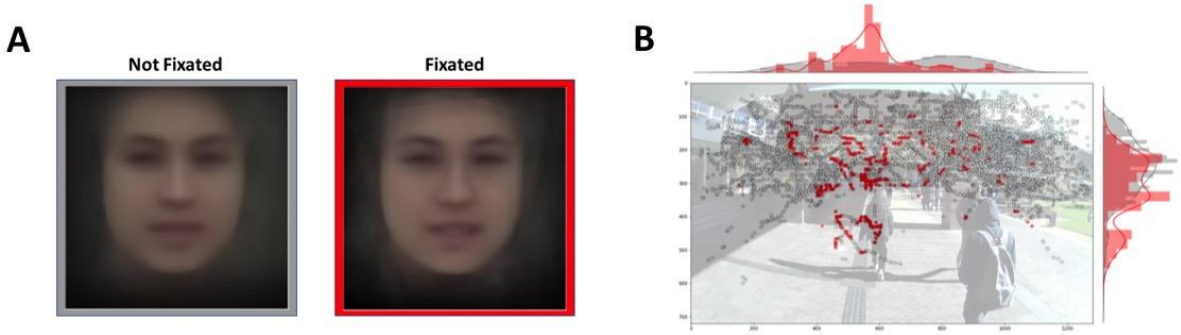


Figure 5.3. Fixated and non-fixated faces detected in the video recordings can reveal the viewer's 'face diet'. (A) Average images of faces that were not fixated (left) and fixated (right) across all participants. (B) Spatial distribution of nose location when faces are detected in the video by the algorithm (grey) and for faces that were fixated by the participant (red). Fixated faces tended to be central in participants' field of view, suggesting that participants mostly directed their attention to people by moving their heads.

Individual differences in naturalistic social attention

Computerised lab-based tests have established that individual differences in social attention are stable across test sessions, and these differences are associated with genetic variation (Constantino et al., 2017). We conducted a correlational analysis of social attention during the navigation task to verify whether stable individual differences in social attention are also found in live natural settings. To test this, we measured the correlation

between individuals' tendency to fixate on people and faces in two distinct segments of the study route that were separated by a short rest break (see Methods).

A scatterplot of participants' proportion of fixations to people in the two route segments is shown in Figure 5.4. We found a significant correlation between proportions of fixations to people by individual participants across the two route segments (Spearman's rho = 0.58, p= 0.001, n = 30), showing the tendency of individual participants to direct attention towards others in natural settings was stable across repeated measurements, at least within this single natural experiment.

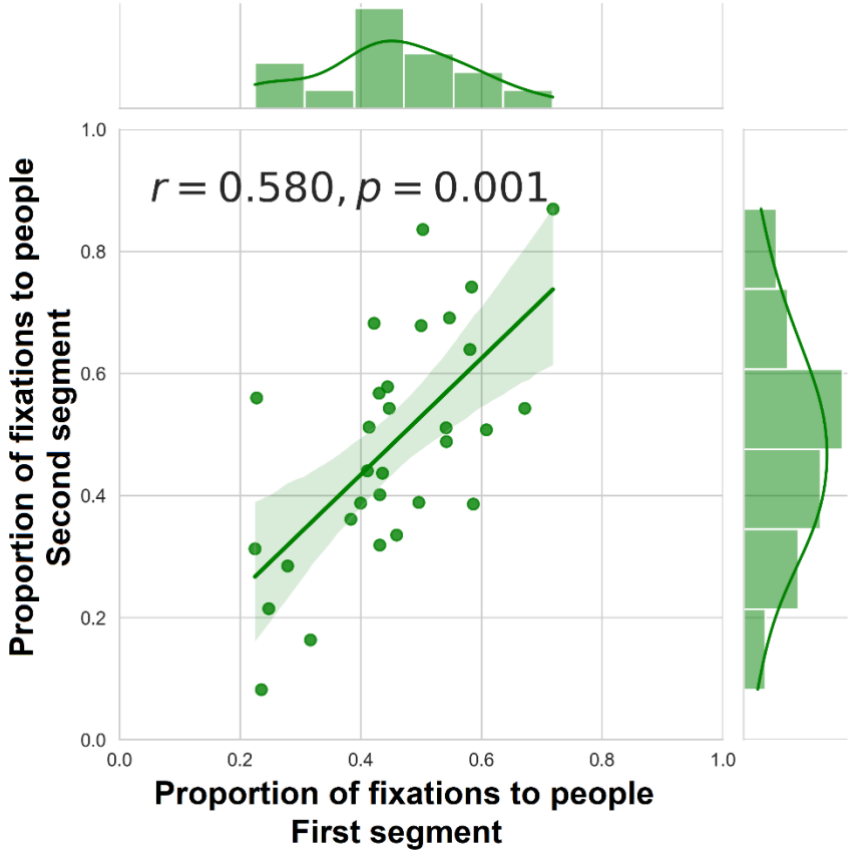


Figure 5.4. Stable individual differences in social attention. Correlation between the proportion of fixations made by individual participants to people in route second segment as a function of the first segment.

Because the university campus was busier for some participants than others, the correlation shown in Figure 5.4 could reflect the relationship between the number of people available in the scene and the participant's proportion of fixations to people. We, therefore,

conducted another correlational analysis, but this time controlling for the influence of the average number of people encountered in video frames. We calculated the residual value for each participant from the linear regression model predicting the probability of fixating people from the average number of people detected in video frames, separately for the two route segments. We found a significant correlation between these residuals for the two route segments (Spearman's rho = 0.532, p= 0.002, n= 30), indicating that some participants tend to fixate more on people than others, regardless of the number of people present (see Supplementary Materials Figure D.2, 'Individual differences analysis of residuals').

Having found reliable individual differences in participants' tendency to fixate on people in their environment, we asked whether these differences were related to face identity processing ability measures. We found no significant correlation between the proportion of fixations to people and their score on an objective (CFMT+: Spearman's rho =-0.166, p=0.371) or self-report measure of face recognition ability (PI-20: Spearman's rho =0.117, p=0.529) (see Methods – Materials). Because previous work has shown that people with high levels of face recognition ability focus more on faces in natural scenes (Bobak et al., 2017), we also repeated this analysis by examining the proportion of fixations to faces only, but again we found no association (CFMT+: Spearman's rho =-0.073, p=0.696; PI-20: Spearman's rho =-0.023, p=0.901).

Fixation patterns during face-to-face interaction associated with face recognition ability

We also recorded participants' fixation patterns during a face-to-face conversation with the experimenter (see Methods – Data collection). This conversation occurred before the main navigation task, as participants listened to scripted task instructions for about 30 seconds before asking any follow-up questions. Because participants were closer to the experimenter, it enabled the face detection algorithm to detect 70 facial landmarks (See Methods). An example of the video recording and the proportion of fixations to each facial landmark for one participant is shown in Figure 5.5A.

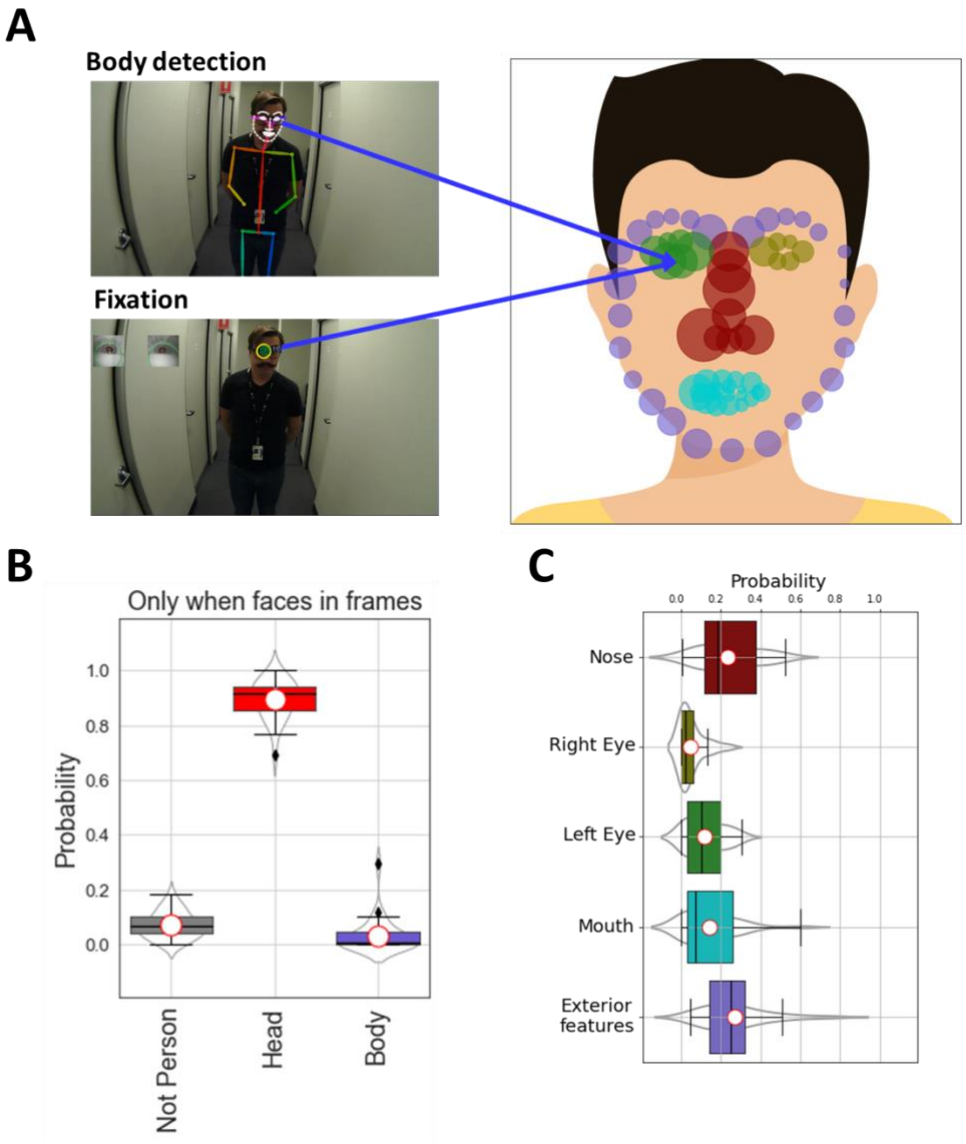


Figure 5.5. Dynamic region of interest analysis of face-to-face interaction. (A) We extracted facial landmarks from the video source using OpenPose (left), and these landmarks were used to register the viewers' fixation positions on the face. The size of circles on the schematic face shows the average proportion of fixations participants made to each landmark (individual participant maps are available in Figure D.3 of Supplementary Materials); (B) Relative frequency of fixations to the experimenter's face and body compared to the surrounding environment; (C) Relative frequency of fixations to facial regions colour coded by landmark in panel A.

Unsurprisingly, participants spent far longer looking at the person when engaged in the conversation compared to the navigation task, with an average of 92.9% of fixations to people and 89.5% to faces (Figure 5.5B). The probability of fixations to different regions of the experimenter's face is shown in Figure 5.5C, showing a focus on internal facial features, in line with screen-based eye-tracking studies (e.g. Yarbus, 1967; Arzipe et al. 2012; Blais et al. 2008). However, these regions are computed by the assignment of fixations to the closest landmark. This method is not precise because landmarks are not distributed evenly

across the face, and so we also used facial landmarks to triangulate precise fixation locations (see Methods – Eye gaze processing). This enabled us to compute heatmaps of participants' gaze patterns, with the average heatmap shown in Figure 5.6A (see Supplementary Materials for individual heatmaps). This average heatmap shows a focus on eyes, nose and mouth in a 'T' shaped distribution, consistent with screen-based eye-tracking studies that almost always find a characteristic of fixations across the internal facial features (e.g. Yarbus, 1967; Arzipe et al. 2012; Blais et al. 2008). Interestingly, there is also a clear leftward bias observable in this figure. This bias is consistent with previous laboratory-based research investigating people looking at faces on screens to perceive identity (Wolff, 1933) and detect emotional expression (Heller & Levy, 1981; David, 1993; Ferber & Murray, 2005).

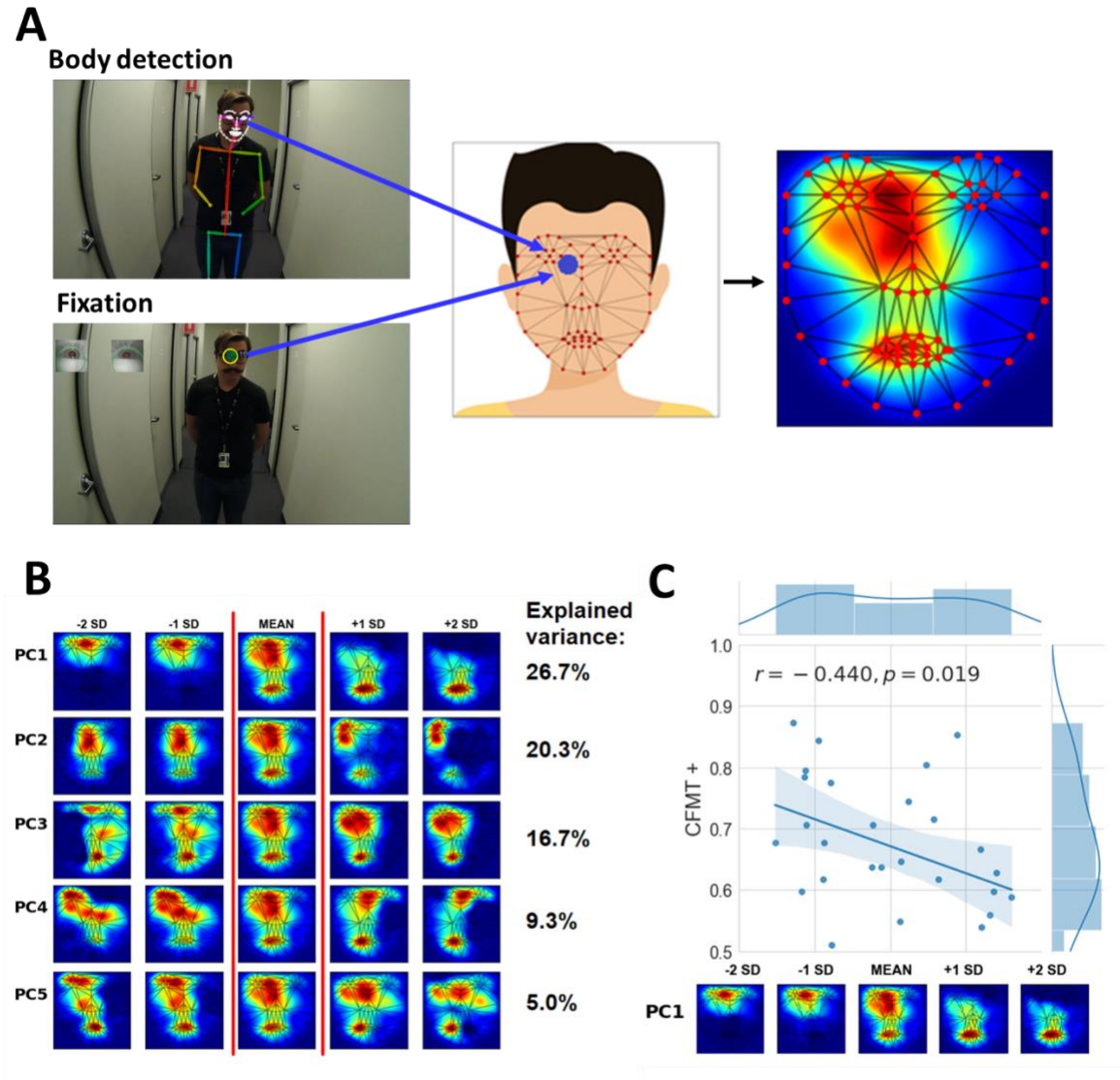


Figure 5.6. Heatmap analysis of face-to-face interaction. (A) Video and eye movement from the wearable eye-tracker (left) were registered using OpenPose facial landmarks and converted to locations on the standard face template using Delaunay triangulation and affine transformations (middle). This technique enabled face fixation data to be aggregated and recorded as heatmaps (right; see Supplementary Materials for individual participant heatmaps). (B) We analysed participant heatmaps using principal component analysis to explore the primary sources of variation in participants' viewing patterns. Here, we show its interaction with the average heatmap in the five first principal components, which accounted for ~78% of the variance. The first principal component explained ~27% of the variance in viewing patterns and captured individual differences in focus on the eye versus mouth regions. (C) Fixations to eyes were associated with higher levels of face recognition ability as measured by the CFMT+.

Previous work shows substantial variation in the gaze patterns of individual participants (e.g. Mehoudar et al., 2014; Arzipe et al., 2017; Dunn et al., 2022). Following our recent eye-tracking work using static eye trackers (Varela et al., 2018; Dunn et al., 2022), we used principal component analysis (PCA) to explore the underlying dimensions of this individual variation in our participant heatmaps (see Methods – Data analysis). Figure 5.6B

visualises the first five principal components in participants' average heatmaps during the face-to-face task. Visual inspection and follow-up analysis showed that the main source of variation (PC1) captured a shift in participants that focussed on the eye region to those that attended to the mouth (Correlation between PC1 and mouth landmark fixation count (n = 28): Spearman's rho = 0.897, $p < 0.001$; Left eye: Spearman's rho = -0.724, $p < 0.001$; Right eye: Spearman's rho = -0.513, $p = 0.005$).

Finally, we measured the association between participants' PC loadings and face recognition ability (CFMT+ score). PC1 showed a significant correlation with face recognition performance (Figure 5.6C, Spearman's rho = -0.44, $p = 0.019$), with no significant correlations with face recognition performance for the other components (PC2 – PC5).

Discussion

Our primary research goal with Experiment 5 was to develop and validate a new research tool to study social attention 'in the wild'. Measures of fixation proportions to people and bodies in a natural setting were broadly consistent with prior research using manual coding of video recordings from wearable eye-trackers (Hessels et al., 2020; DeLillo et al., 2021), and there was high agreement between our automated measures and manual experimenter coding (see Figure D.7). We found indexes of social attention to be reliable over repeated measurements. Together, we interpret this as evidence that dynamic region of interest (dROI) approaches are valid for studying social attention in natural settings.

The dROI approach enabled us to ask some preliminary questions inspired by screen-based studies of social attention. We first examined the extent to which faces automatically capture attention as participants navigated a busy public space. Contrary to conclusions based on lab-based experiments (e.g. Vuilleumier, 2000; Bindemann et al., 2005; Theeuwes & Van der Stigchel, 2006; Gamer & Büchel, 2009; Ro, Russell & Lavie, 2001; Bobak et al., 2017; Rösler, End & Gamer, 2017; Gregory, Bolderston & Antolin, 2019), we found very little evidence that faces capture attention in this situation. Fixations to faces – when faces were visible in the participants' field of view – made up a small proportion of total fixations (14%). Moreover, when comparing attention capture by faces and bodies that were fully visible and those that were only partially viewable, we found that this increased fixations to both

faces and bodies equivalently. This evidence does not support the idea that people automatically orient their attention to faces, at least for unfamiliar faces in a public space.

As expected, we found that participants spent a far greater proportion of time looking at faces during one-to-one social interaction. Fixation patterns to faces are known to be highly context-dependent, for example, dependent on whether the face is moving (Buchan et al., 2007; Foulsham et al., 2010; Scott, Batten & Kuhn, 2019), speaking (Buchan et al., 2008; Võ et al., 2012) and the non-verbal behaviour of the viewed person (Võ et al., 2012; see Hessels, 2020 for a review). This evidence suggests that the role of context is more important than the intrinsic properties of faces themselves in determining attention to faces, underlining the complexity of our visual experience with faces in daily life.

Using automated face detection combined with eye-movement data enabled us to visualise the faces that people fixated on in our study, offering a window into participants' perceptual experience of faces. Limits of the resolution of the video frame meant that we were only able to visualise a subset of the viewed faces, but we anticipate that improvements in wearable video cameras and eye-tracking technology can allow richer analysis in future work using this approach. In general, it is surprising how little information is available about the amount and quality of perceptual experience people have in a typical day. This information is important theoretically because our exposure to faces is the basis of our expertise in face processing relative to other classes of objects (see Young & Burton, 2018). But quantitative and qualitative analyses of this exposure are very rare. Studies using head-worn cameras focus exclusively on infancy and childhood experience (e.g. Jayaraman et al., 2015; Fausey et al., 2016; Sugden & Moulson, 2019) – despite abilities in face recognition continuing to develop up to people's mid-thirties (Germine et al. 2011; Dunn et al. 2020) – and these studies have not examined participants viewing behaviour. Future studies using the dROI approach can better understand how viewers sample visual information from people in their everyday lives.

The dROI approach also has significant potential for understanding individual differences in social attention and face-processing ability. We found that individual differences in attention to people were stable across different study sections, consistent with studies using screen-based approaches where there appears to be a genetic basis to people's social attention (Constantino et al., 2017; Kennedy et al., 2017). We also found that patterns of fixations in a face-to-face interaction were associated with face recognition

ability, adding to a growing list of studies that have found associations between face processing ability and face information sampling patterns (e.g. Dunn et al., 2022; Wilcockson et al., 2020; Varela et al., 2018; Bobak et al., 2017; Bird el al., 2011; Bal et al., 2010; Riby et al., 2009; see Avidan & Behrmann, 2021 for a review). Across these studies, fixation patterns associated with high performance are not consistent. For example, we found that people with higher face recognition ability tended to make more fixations on the eyes, but while some screen-based studies found this pattern (Wilcockson et al., 2020; see also Tardif et al., 2019), others report no association (e.g. Arzipe et al. 2017), and others found the opposite result (e.g. Dunn et al., 2022). Again, one explanation of this inconsistent picture may be the important role of context in the information sampled from a face, pointing to the need for research that examines the relationship between face processing, social attention and social cognition in more diverse settings and situations.

We are hopeful that the approach we have presented here can facilitate this type of naturalistic social perception research. The present study scarcely scratches the surface of what is possible. Future work could take on ambitious aims, for example, to capture a more complete picture of people's perceptual exposure to faces in their daily lives. Intuitively, this exposure contains rich diversity, such as the familiarity of people we encounter, the contexts and viewing conditions we encounter them in, and the nature of our social interactions. Characterising the multidimensional nature of this perceptual data and differences in how individuals sample it should be critical information to underpin the development of theory in this field.

Chapter 6 - General Discussion

Summary of research aims and findings

This thesis investigated the strengths and weaknesses of humans and DCNNs performing face identification tasks. Unfamiliar face identification is a critical task in forensic scenarios, but achieving accurate performance is still challenging for both humans and DCNNs. Consequently, researchers have been developing mechanisms to improve facial identification decisions. So far, researchers suggest that using humans with high performance in facial identification decisions, state-of-the-art DCNNs, or even a combination of them might improve performance in facial identification decisions (e.g. Phillips et al., 2018). However, understanding the differences in the usability of these alternatives still needs to be clarified. And so, in this thesis, we investigated the potential use of humans and DCNNs performing such tasks and the different strengths and weaknesses they engage when coding facial identity information. We employed innovative tools using engineering approaches across seven studies reported in four experimental Chapters. In these studies, we found evidence that humans and DCNNs employ different strategies to achieve unfamiliar facial identification decisions. Interestingly, we also show that the significant differences within/between humans and DCNNs can be crucial for optimising identity verification.

Chapter 2 sought to investigate if image quality in facial identification affected humans and DCNNs differently. As shown, security services often perform the task of identity comparison using -for example- a reference against images of the offender. However, such images may be of poor quality - with distortions and bad angles of view (e.g. taken from CCTV cameras). And so, in Chapter 2, we created an unfamiliar face-matching task aiming to understand how typical viewers (i.e. students) and 9 DCNNs would be affected by changes in image quality. This task contained a total of 50 trials, each containing two frontal face images. Humans and DCNNs were, required to decide whether the images were of the same person. They processed the stimuli under two conditions: Same-Resolution and Different-Resolution. In the Same-Resolution condition, the two face images were similarly degraded by filtering (i.e. both had the 'same' resolution). In the Different-Resolution condition, one image remained in perfect condition. In addition, we used two distinct low-pass filtering paradigms for manipulating image quality: Fast Fourier

Transformation and Gaussian filtering. Our results show a significant two-way interaction between humans and DCNNs performing the stimuli conditions. That is, DCNNs were more affected by the Different-Resolution condition compared to humans. This important interaction indicates that humans and DCNNs process identity information differently.

Chapter 3 then investigated the basis of humans and DCNNs being used in a human-AI teaming. The investigation led by us in the previous chapter showed that humans and DCNNs suffer considerably when processing facial identity verification when the source images are of poor quality. However, interestingly, previous research suggested that combining their independent decisions in a single Human-AI 'fused' response could potentially improve their overall performance in the task. Therefore, our objective was to investigate the underpinnings for such a combination to enhance unfamiliar facial identification decisions. For that, we examined multiple Human-AI setups to understand how fusing independent decisions made by humans and DCNNs can lead to improvements in identity verification. This investigation showed that improvements caused by fusing humans and DCNNs processing face identification decisions have a robust linear relationship with the accuracy difference between them. To illustrate, our results show that humans need to perform no worse than 10% lower than the DCNNs' performance to show improvements in their fused identity verification decisions compared to DCNNs alone. In addition, the linear model suggests that humans with similar performance compared to DCNNs', can improve the DCNN's performance by 5%. Interestingly, we show an additional accuracy increase proportionally related to disagreements in humans-DCNN decisions when performing face identification decisions. Ultimately, the optimal human-AI team comprises a human who performs at a similar level of accuracy to the DCNNs in facial identification decisions but ranks similarities differently than the DCNN.

In Chapter 4, we aimed to understand how information sampling would predict individual differences in face-processing abilities. Past research modelling human face processing suggests that humans process faces 'holistically' rather than feature-by-feature (e.g. Farah, Wilson, Drain & Tanaka, 1998; Richler & Gauthier, 2014). Studies aiming to measure if individual differences in face processing are somewhat linked to improved holistic processing found weak to no correlation (e.g. Richler, Cheung, & Gauthier, 2011; Rezlescu, Susilo, Wilmer, Caramazza, 2017), while other studies found some association of individual differences with part-based processing (Sunday, Richler, & Gauthier, 2017). Such

mixed results show that the mechanisms that drive individual differences in face-processing ability remain unclear. And so, in Chapter 4, we developed computational methods for more direct investigation regarding the mechanisms leading to improved face processing ability. We investigated differences in information sampling of facial features that human participants engaged when matching and recognising faces. Our approach allowed the visualisation of eye-tracking data that described individual differences in face-matching and face-recognition abilities. We found that high-performers generally explored more facial regions than typical viewers when coding identity information, contrary to previous eye-tracking studies showing that super-recogniser process faces holistically (e.g. Bobak et al., 2017). Interestingly, a subsequent approach measuring the quality of identity information using DCNNs revealed that super-recognisers extract more high-value identity information from faces.

In Chapter 5, our aim was to verify if the mechanisms of face processing found in laboratory screen-based stimuli were similar to the ones found in the wild. As we have shown throughout this thesis, most of the studies of face perception which investigate attention using eye trackers are screen-based studies (e.g. Yarbus, 1967; Amso, Haas & Markant, 2014; Birminghan, Bischof & Kingstone, 2009; Bobak et al., 2017; Rösler, End & Gamer, 2017; Gregory, Bolderston & Antolin, 2019). However, investigating attention to static images does not necessarily represent the dynamic and multidimensional reality of our social experience, as contrived laboratory tasks are inadequate analogues of real-world when investigating social attention (Kingstone, 2009; see also Nasiopoulos, Risko & Kingstone, 2015; Risko, Richardson & Kingstone, 2016). In addition, as we have illustrated, analysing attention 'in the wild' is not an easy task because the output of mobile eye trackers is rich in data information and therefore requires very labour-intensive manual coding of fixation positions to dynamic areas of interest. And so, in Chapter 5, we aimed to facilitate subsequent studies of social attention in the wild by developing a fully automatic methodological approach that enables fine-grained investigations of mobile eye-tracking data. By using - and validating - our approach in our study, we show strong evidence that static screen-based laboratory tasks do not reflect the mechanisms of attention found in the wild. As an example, contrary to conclusions based on screen-based experiments (e.g. Vuilleumier, 2000; Bindemann et al., 2005; Theeuwes & Van der Stigchel, 2006; Gamer & Büchel, 2009; Ro, Russell & Lavie, 2001; Bobak et al., 2017; Rösler, End & Gamer, 2017;

Gregory, Bolderston & Antolin, 2019), we found very little evidence that faces capture attention when participants navigated public spaces. Interestingly, when engaging in face-to-face interaction, participants spent a greater proportion of time looking at faces compared to when navigating the public space. Such results suggest that context is far more significant than the intrinsic properties of faces themselves in determining attention toward faces.

Main Findings

Face processing tasks are difficult for humans and DCNNs

Our results reiterate the common finding that unfamiliar face matching is challenging. Both human participants and DCNNs performed a series of unfamiliar face-matching and face-recognition tasks. Starting from the theoretical perspective that people are 'experts' in processing the signals emitted by faces (e.g. Carey, 1992), our results are consistent with the extensive literature suggesting that this expertise is reserved only for familiar faces (e.g. Young & Burton, 2018). In this thesis, we simulated potential real-world situations where humans and DCNNs required to identify individuals depicted in images. We found that both humans and DCNNs can potentially misinterpret the identity signals emitted by faces. And so, given the importance of the task in applied settings, this misinterpretation is problematic as it could result in – for example- the arrest of an innocent person.

Humans and DCNNs possess different strategies for processing identity information

In Experiment 1, we showed that humans and DCNNs could suffer decrements in performance when processing unfamiliar face-matching tasks when images were of poor quality. Notably, this reduction in performance when processing images of poor quality can lead to profound outcomes regarding face identification in applied settings. In such settings, it is often common to compare mugshot images of suspects with images from -for example- CCTV footage. Importantly, as seen, such images can contain distortions and be presented in poor quality (see Seckiner et al., 2018).

Importantly, however, our investigation revealed that DCNNs suffered substantially more than humans when image pairs were of different qualities compared to when pairs were of similar quality. This implies that humans and DCNNs employ different strategies to achieve broadly similar levels of performance in face identification tasks. In terms of identity information contained within the images, our results indicate that humans possess a superior ability compared to DCNNs to extract the remaining configural information from poor-quality faces and compare this information to high-quality images. In addition, we tested such interaction using different low-pass filters to degrade the featural information of the images - and found a similar pattern of results but in different magnitudes. This evidence of robustness in our results is critical because it further suggests that the differences found between humans and DCNN processing images of varying quality will generalise to a variety of real-world scenarios.

Such differences in the cognitive process of humans and DCNNs processing faces can be advantageous when working together to improve the quality of the decision regarding someone's identity. Previous work has found that combining decisions made by DCNNs and humans can improve identity verification accuracy (e.g. Philips et al., 2018). More, increased diversity among decision-makers, when combined, further improves facial identity verification performance (White et al., 2013; Jeckeln et al., 2018). The fact that we have found processing differences with respect to how DCNNs are able to match faces differing in image quality could point to processing differences that could be exploited in future work. Because of this apparent difference in approach, teaming between humans and DCCNs should result in superior performance compared to either humans or DCNNs alone (see also Towler et al., under review; Hong & Page, 2004). From a theoretical perspective, this result also contradicts the idea that DCNNs could potentially be used to model face processing in humans (e.g. see Jozwik et al., 2022).

Humans and DCNNs used in conjunction improve the quality of face verification decisions

We combined responses made by humans and DCNNs processing unfamiliar identity verifications and found that their combined decisions can be more accurate than those made by either humans or DCNNs alone. It is increasingly common in applied settings to use facial identification software as a solution to overcome error-prone human decisions in face

identity verification (e.g. White et al., 2015, Grother et al., 2019). However, as demonstrated in this thesis, using only DCNNs as a tool for identity verification tasks may not be optimal. As an alternative, studies have implemented combining (i.e. fusing) decisions made by humans and DCNNs and found significant performance improvements in facial identity verification (e.g. Phillips et al., 2018; Towler et al., under review). Still, despite showing possible improvements in face identification performance by fusing humans and DCNNs, the basis of this advantage is unclear. Understanding how humans can increase the accuracy of DCNNs' decisions in identity verification is significant because it shows that humans remain a valuable tool for such tasks in applied settings.

We investigated multiple Human+DCNN scenarios to provide a clearer picture of the mechanisms behind their fusion and strategies in which they can provoke improvements in facial identification decisions. Such an extensive investigation demonstrates potential solutions for using humans to improve DCNNs' decisions performing facial identification decisions. We found that combining decisions made by humans with DCNNs of similar performance caused significant improvements in the overall quality of the face identification decisions. In addition, we found even further improvements when combining the responses of humans who somewhat disagree with their DCNN pair. Our results expand previous studies investigating fusion effects in humans+DCNNs because it gives a practical guide for when teaming will be beneficial in applied settings.

Information sampling provides routes for expertise in face-processing tasks

Information sampling explains individual differences in identity processing ability. So far, this thesis suggests that humans of superior performance can usefully contribute to the decision-making of DCNNs. It is well documented in the literature that human ability in identity processing is responsible for large individual differences (e.g. Dunn et al., 2020, Burton et al., 2010, Russell et al., 2009). Some studies suggest that a more holistic perception of facial features results in improved face processing abilities (e.g. Bobak et al., 2017; Bennetts, Mole & Bate, 2017). However, such a result is unclear because: (i) in their studies, participants passively viewed images of faces instead of specifically being asked to code identity information; (ii) studies found that measures of holistic measures do not necessarily correlate with face processing abilities (e.g. Rezlescu et al., 2017).

To provide a clearer picture regarding individual differences in face identity processing, we developed innovative methodological approaches to directly investigate the mechanisms of attention of humans while processing face-matching and recognition tasks. We developed an approach to analyse and visualise the rich data from eye-tracking devices while our participants engaged in the different face identity verification tasks. As predicted, our approach illustrated significant information sampling differences between - and within – humans performing the stimuli, revealing critical information sampling characteristics that explained individual differences in identity processing.

Overall, our results point out that improved face processing abilities relate to more exploration of facial features, directly contradicting the global 'holistic' idea of face processing in superior performance. In addition, to further test if holistic processing relates to face processing abilities, our gaze-contingent condition revealed that obstructing foveal information reduced overall performance. Still, those with superior performance maintained their status throughout all aperture conditions. Ultimately, our results suggest that improved performance is related to the improved exploration of facial features, showing a potential higher motivation and efficiency in super recognisers to use featural information to achieve superior identity recognition performance throughout all stimuli aperture conditions.

Human-guided information sampling benefits DCNNs for facial identity information

Throughout our experimental designs, we found that superior facial identity recognition performance relates to the improved exploration of facial features. Such a result suggests that super-recognisers employ multiple high-quality sources of featural information to potentially create a more robust representation of identities for subsequent recognition. However, to our knowledge, no other studies compared the quality of face identity information sampling between humans of varied facial identity recognition abilities. And so, it was unclear if super-recognisers' information sampling would reflect more identity information than typical viewers. To investigate this, we developed a methodological approach using mathematically designed retinal filters (Targino Da Costa & Do, 2014) to create static face images representing the human participants' perceptual

sampling when coding for facial identity information. DCNNs then compared such images with others from the same or different identities to measure the quality of identity information sampled by human participants. Notably, as a reminder, we also created a condition containing randomised fixation positions for a more solid investigation.

In general, our human/random facial information sampling investigation revealed that DCNNs could more easily detect identity information when using human-guided sampling. When processing a facial verification procedure using human-guided attention, DCNNs showed a significant performance improvement compared to the randomised condition. This result then indicates that the information humans judge to be critical for identity verification potentially corresponds to the same information DCNNs consider critical for coding identity information. And so, despite our previous results showing different strategies between humans and DCNNs when comparing images of different image qualities, there is an observable overlap in the featural information in high-quality images that they use for processing identity information (see Lai et al. 2020; Rong et al. 2021; Yang et al. 2022). In addition, our results show that the information sampled by super-recognisers produced significantly higher performance compared to typical viewers' and randomised information. This improvement in the performance of DCNNs using super-recognisers' perceptual sampling shows that higher facial identity verification ability relates to exploring not only a significantly higher number of facial features but also qualitative information resulting in a more robust representation of identities for subsequent verification.

Social attention in the wild conflicts with screen-based research

As shown throughout this thesis, studies of adults' attention to people in natural settings are extremely rare, and almost all knowledge on this topic comes from tightly controlled laboratory-based research. This laboratory-based research shows, for example, that faces capture attention and are processed preferentially relative to non-face objects and bodies (e.g. Bindemann et al., 2005), supporting the view that face processing is automatic (Palermo & Rhodes, 2007). However, it was not known whether this would hold for ambient environments populated with many competing stimuli – each with its unique affordance (Gibson, 1979) – and where the 'social stimuli' are real people instead of manipulated photographs on computer screens. To test this, we developed a unique

method to investigate social attention in 'the wild' by studying the eye movements of participants wearing eye-tracking devices which monitor their fixations as they navigate real-world ambient environments.

In this thesis, we introduced a novel method that automates fine-grained investigations of naturalistic social attention for the first time. Our 'dynamic regions of interest' (dROI) approach automatically measures social attention in ambient environments frame-by-frame. We achieved this by co-registering eye-movement data from a wearable eye-tracker with body and face landmark positions extracted from video data using a state-of-the-art computer vision algorithm (Cao et al., 2019). This encodes eye fixations directed towards people and maps fixations to landmarks on the face and body. Our approach overcame many significant limitations of prior work on social attention in natural settings, saving substantial research effort by avoiding the need for manual coding of fixations to pre-specified regions. In addition to removing the burden of manual coding, our approach also increased temporal resolution and the volume of data, enabling new analytic approaches which open up new avenues to study person perception 'in the wild'.

Our methodological approach enabled us to ask some preliminary questions inspired by screen-based studies of social attention in natural settings. We examined the extent to which faces automatically captured attention as participants navigated a busy public space. Contrary to conclusions based on lab-based experiments (e.g. Bobak et al., 2017; Rösler, End, & Gamer, 2017; Gregory, Bolderston, & Antolin, 2019), we found no evidence that faces captured attention 'in the wild'. Curiously, we show that fixations to faces – when faces were visible in the participant's field of view – made up a small proportion of total fixations. Moreover, when comparing attention capture by faces - and bodies - that were fully visible and those that were only partially viewable, we found that this increased fixations to both faces and bodies equivalently. This evidence does not support the idea that people automatically orient their attention to faces, at least for unfamiliar faces in a public space. Still, our study also points to two additional directions for future research made possible by our methodological dROI approach. First, it reveals new possibilities for understanding individual differences in social attention and face-processing ability. Individual differences in attention to people were stable across different sections of the navigation route, consistent with screen-based eye-tracking studies that show -for example- a robust hereditary influence on patterns of attention to social scenes (e.g. Constantino et

al., 2017; Kennedy et al., 2017). Although the sample size in our study was not large enough to make strong inferences about whether these individual differences transfer to naturalistic settings, they still provided some assurance of measurement reliability at the individual level.

In contrast, fixation patterns when engaging in a face-to-face interaction predicted face recognition ability. Interestingly, despite this task being a real-world interaction, our results corroborate with screen-based task studies' results. That is, we found that participants prioritised engaging in observing faces significantly more than other regions of the scene, and -most interestingly- their fixation patterns attending to facial regions predicted individual differences in face-processing ability. This result is interesting because it reveals potential routes for developing new theories regarding the underpinning of superior face recognition. For example, a larger proportion of fixations observing the eye region during the conversation predicted face recognition performance. It is known that the eyes convey cues for mental state. And so, perhaps individuals with higher facial recognition performance also seek out cues to the mental state of the subject. These are potential routes that future studies could address to robust our understanding regarding superior face processing abilities and individual differences reflected by fixation patterns in the wild.

Practical Implications

Unfamiliar face identity verification is a critical task performed by applied security settings by humans and facial recognition technology. When performed by humans, this thesis shows large individual differences and proportions of errors in the decisions regarding identities depicted in images. This is consistent with previous work that shows a similar outcome in professionals who perform the task daily (see also White, Towler, & Kemp, 2021). Thus, an interesting approach for overcoming humans' error-prone unfamiliar facial identification ability is to automatise decisions by using facial recognition technology. The multi-layer architecture of such technology, such as the ones found in Deep Convolutional Neural Networks (DCNNs), means that artificial systems can now achieve accurate face identification across a wide range of image variations. However, as shown, such technology is also error-prone when performing unfamiliar identity verification. In particular, prior research has identified ‘blindspots’ in the algorithm training, causing, for example, poorer

performance for certain demographics compared to others (e.g., ethnicity: see Cavazos, Phillips, Castillo, & O'Toole, 2020).

In my thesis, I also found that algorithms have a 'blindspot' when matching images that vary in image quality (see also Vera-Rodriguez et al., 2019). Similarly to the differential accuracy found for different demographic groups (Grother et al., 2019), this is likely to be due to the composition of image sets used in DCNN training. Prior work has shown that training algorithms with blurred images can produce more robust performance when later matching blurred images (Vogelsang et al., 2018), and so one potential solution to this problem is to introduce blurred images into the algorithm training. This would be a worthwhile investigation in future work, given the practical importance of this type of low-to-high quality image matching in applied settings.

Another solution that this thesis has shown to be promising is to combine face identity decisions made by humans and DCNNs. Prior work has investigated the benefit of 'fusing' independent judgments made by humans and DCNNs. Such studies found potential improvements – even reaching ceiling levels - in facial identification performance (e.g. Phillips et al., 2018; Towler et al., under review). In this thesis, we investigated the outcomes of such fusion further and found a model that explains the fusion effect of humans and DCNNs performing independent decisions regarding identity verification. Our results illustrate profound implications in applied settings because it provides guidelines regarding how and when to team humans and DCNNs to significantly reduce errors in facial identity verification. Given how critical the task is, we show that combining the decisions of humans and state-of-the-art DCNNs could potentially increase the quality of identity verification decisions compared to humans or DCNNs alone.

The humans and DCNN components of a team must demonstrate similar unfamiliar identity processing abilities if we want the team to outperform either component alone. We investigated the effects of combining the decisions of humans with similar accuracy to DCNNs. We found substantial improvements in facial verification performance when both have similar accuracy - measured by the same test. Throughout this thesis, we showed pieces of evidence suggesting that it is necessary to select individuals based on face identification ability (i.e. super-recognisers) as potential solutions for performing unfamiliar facial identification procedures (e.g. Davis, Forrest, Treml, & Jansari, 2018; Davis, Lander, Evans, & Jansari, 2016; Robertson, Noyes, Dowsett, Jenkins,& Burton, 2016). However, our

results provoke profound implications because it provides a new benchmark to determine the baselines of human performance to work in applied settings alongside DCNNs. For example, instead of selecting individuals based on a threshold in the distribution of human performance on a given face processing test (e.g. the top 2% performers on the CFMT+: Russell et al., 2009), our results argue that the DCNNs' ability on the given practical test should be what determines the threshold for selecting humans. For example, we might require that humans are within 10% of the performance of the DCNN in order to start contributing to a teaming approach. More, it is important to address that the accuracy of facial recognition technology is evolving at rapid paces, and so applied settings should constantly check the unfamiliar face processing abilities of their staff for accurate fusion and decisions.

Furthermore, researchers show that working under time pressure (Fysh & Bindemann, 2017), sleep restriction (Beattie, Walsh, McLaren, Biello, & White, 2016), and high anxiety levels (Attwood, Penton-Voak, Burton & Munafò, 2013) can significantly affect the outcome of human facial identification decisions. Thus, to properly work alongside a state-of-the-art DCNN, its human sidekick should always be at its peak performance (i.e. similar to DCNNs') for accurate facial verification decisions. Failure to ensure this will ultimately result in teams in which the human negatively impacts performance – contributing to errors rather than enhancing performance.

Methodological Implications

This thesis investigated individual differences in humans processing a series of studies containing unfamiliar face identification. In some of our studies, eye-tracking devices were essential tools for achieving our research questions. However, such devices produce rich and complex data for analysis and visualisation. And so, our approach for analysing and visualising such data was due to the use of innovative paradigms using, for example, statistical learning (i.e. the PCA from Chapters 4 and 5) and machine learning tools (i.e. the dROI approach from Chapter 5). Not only, we also offer evidence that such new approaches would result in similar outcomes compared to the traditional methods for analysis, but with easier use for analysis and visualisation of results. And so, the methods developed here will be of great use for future research because it allows much greater realism in the

environments in which we test face recognition and will also allow easy replicability and usability for face perception moving forward.

Theoretical implications

Attention to faces is context-dependent. Throughout this thesis, we investigated humans performing a series of tasks involving faces, such as face-matching and recognition tasks, navigating ambient environments, and engaging in face-to-face conversations. In such studies, our objectives were to investigate individual differences and the processes reflecting superior face-processing performance. Contrary to previous research suggesting that superior performance is reflected by the greater use of holistic facial information (e.g. Bobak et al., 2017), our results argue that performance in face-processing tasks relates to the exploration of facial features. However, a deeper investigation of such results revealed that each task produced a different outcome which explains superior facial processing performance.

For example, on the one hand, in Chapter 4, we show that superior performance in face-matching tasks is explained by greater use of the eye region. On the other hand, superior performance in face-recognition tasks is explained by avoiding the eye region. Notably, these apparently contradictory results can be reconciled because the tasks are different. And so, our results argue that the critical role of context is at least as important as the intrinsic properties of faces themselves in determining facial features reflecting superior performance. In addition, we provide evidence that screen-based studies do not necessarily reflect the patterns of face processing found in the wild. Previous studies have already encountered that fixation patterns change when faces move (Buchan et al., 2007; Foulsham et al., 2010; Scott, Batten & Kuhn, 2019), speak (Buchan et al., 2008; Võ et al., 2012) and engage on non-verbal behaviour (Võ et al., 2012; see Hessels, 2020 for a review). Here, we expand our knowledge by showing that attention to faces reflecting individual differences is also flexible to the nature of the task.

The flexibility in the attention of humans processing faces in divergent tasks shows a higher complexity for facial recognition technology to model human face processing. Given the incredible facial identity verification performance of facial recognition algorithms and their robustness across a wide range of image variations, an emerging body of research

suggests that DCNNs could be used as a model for comprehending the basis of face processing in humans (e.g. Dobs et al., 2022; Grossman et al., 2019; O'Toole et al., 2018; Kuzovkin et al., 2018; Ratan Murty et al., 2021; Tsantani et al., 2021). Similar to Chapter 2 of this thesis, these studies compare humans and DCNNs engaging in a face-processing-related task to infer commonalities or divergences in their responses to stimuli. However, the flexibility of humans in actively attending different regions of interest depending on the stimuli reveals that the face-processing model cannot be described as a single generalised model. Instead, it illustrates that the mechanisms of face processing in humans are highly complex and task-dependent. And so, despite DCNNs being as effective as humans – for example- performing identity recognition (Grother et al., 2019), our results demonstrate that similarities in their processes in the tasks arguably do not code for similar cognitive mechanisms.

Conclusion

This thesis contributes to understanding unfamiliar identity processing in humans and DCNNs. From this body of studies, we identified several factors resulting in expanding our current knowledge regarding individual differences in face-processing abilities. Analysing our results using innovative engineering approaches, we identified that humans possess a more flexible capacity for processing separable configural and featural facial identity information than DCNNs. Interestingly, despite being error-prone for identity processing, humans and DCNNs are plausible solutions to work together to perform the task in forensic settings. In addition, we show that individual differences in face processing can be explained by the information sampled when engaging in the task. However, we also show that the results found in the wild do not reflect patterns of results in contrived screen-based stimuli. Ultimately, the findings of this thesis not only improve our theoretical understanding of the nature of face identity perception but also show that multidisciplinary methods can be developed - and used - in future research to explore face perception in humans.

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3), 40-40.

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition*, *182*, 73-83.

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 997.

Amso, D., Haas, S., & Markant, J. (2014). An eye-tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PloS one*, *9*(1), e85701.

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.

Arizpe, J., Walsh, V., Yovel, G., & Baker, C. I. (2017). The categories, frequencies, and stability of idiosyncratic eye-movement patterns to faces. *Vision Research*, *141*, 191-203.

Attwood, A. S., Penton-Voak, I. S., Burton, A. M., & Munafò, M. R. (2013). Acute anxiety impairs accuracy in identifying photographed faces. *Psychological Science*, *24*(8), 1591-1594.

Avidan, G., & Behrmann, M. (2021). Spatial integration in normal face processing and its breakdown in congenital prosopagnosia. *Annual Review of Vision Science*, *7*, 301-321.

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity?. *European Journal of Cognitive Psychology*, *3*(1), 87-103.

Bal, E., Harden, E., Lamb, D., Van Hecke, A. V., Denver, J. W., & Porges, S. W. (2010). Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*, *40*(3), 358-370.

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*(1), 1-13.

Bartlett, J. C., & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive psychology*, *25*(3), 281-316.

Bartlett, J. C., Searcy, J., H., Abdi, H., (2003). "What are the routes to face recognition?", in Perception of Faces, Object and Scenes: Analytic and Holistic Processes Eds M A Peterson, G Rhodes (New York: Oxford University Press) pp 21 ^ 52

Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: A critical review and future directions. *Frontiers in Human Neuroscience, 8*(491).

Bate, S., Haslam, C., Tree, J. J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*(7), 806-819.

Bate, S., Parris, B., Haslam, C., & Kay, J. (2010). Socio-emotional functioning and face recognition ability in the normal population. *Personality and Individual Differences*, *48*(2), 239-242.

Beattie, L., Walsh, D., McLaren, J., Biello, S. M., & White, D. (2016). Perceptual impairment in face identification with poor sleep. *Royal Society open science*, *3*(10), 160321.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: face-blind from birth. *Trends in cognitive sciences*, *9*(4), 180-187.

Benjamins, J. S., Hessels, R. S., & Hooge, I. T. (2018, June). GazeCode: Open-source software for manual mapping of mobile eye-tracking data. In *Proceedings of the 2018 ACM symposium on eye-tracking research & applications* (pp. 1-4).

Bennetts, R. J., Mole, J., & Bate, S. (2017). Super-recognition in development: A case study of an adolescent with extraordinary face recognition skills. *Cognitive neuropsychology, 34*(6), 357-376.

Best-Rowden, L., Bisht, S., Klontz, J. C., & Jain, A. K. (2014, September). Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *IEEE International Joint Conference on Biometrics* (pp. 1-8). IEEE.

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision research*, *50*(23), 2577-2587.

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology*, *27*(6), 707-717.

Bindemann, M., Avetisyan, M., & Blackwell, K. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*, 378–386.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognise unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied, 18*, 277–291.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition, 1*, 96–103

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & De Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12*(6), 1048-1053

Bindemann, M., Fysh, M. C., Trifonova, I. V., Allen, J., McCall, C., & Burton, A. M. (2021). Face Identification in the Laboratory and in Virtual Worlds. *Journal of Applied Research in Memory and Cognition*.

Bird, G., Press, C., & Richardson, D. C. (2011). The role of alexithymia in reduced eye-fixation in autism spectrum conditions. *Journal of Autism and Developmental Disorders*, *41*(11), 1556-1564.

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2).

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Social attention and real-world scenes: The roles of action, competition and social content. *The Quarterly Journal of Experimental Psychology*, *61*(7), 986-998.

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PloS one*, *3*(8), e3022.

Blanton, A., Allen, K. C., Miller, T., Kalka, N. D., & Jain, A. K. (2016). A comparison of human and automated face verification accuracy on unconstrained image sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 161-168).

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2020). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, 104341.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A. & Bate, S. An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex, 82*, 48–62 (2016).

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one*, *11*(2), e0148148.

Bobak, A. K., Jones, A. L., Hilker, Z., Mestry, N., Bate, S., & Hancock, P. J. (2023). Data-driven studies in face identity processing rely on the quality of the tests and data sets. *Cortex*, *166*, 348-364.

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *Quarterly Journal of Experimental Psychology*, *70*(2), 201-217.

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

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.

Brunsdon, R., Coltheart, M., Nickels, L., & Joy, P. (2006). Developmental prosopagnosia: A case analysis and treatment study. *Cognitive Neuropsychology*, *23*(6), 822-840.

Bruyer, R., & Coget, M. C. (1987). Features of laterally displayed faces: Saliency or top-down processing?. *Acta Psychologica*, *66*(2), 103-114.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1-13.

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162-171.

Bülthoff, I., Mohler, B. J., & Thornton, I. M. (2019). Face recognition of full-bodied avatars by active observers in a virtual environment. *Vision Research*, *157*, 242-251.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*, 943–958

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256–284.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior research methods*, *42*(1), 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248.

Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia*, *48*(14), 4057-4092.

Butcher, N., & Lander, K. (2017). Exploring the motion advantage: evaluating the contribution of familiarity and differences in facial motion. *Quarterly Journal of Experimental Psychology*, *70*(5), 919-929.

Butcher, N., Lander, K., Fang, H., & Costen, N. (2011). The effect of motion at encoding and retrieval for same-and other-race face recognition. *British Journal of Psychology*, *102*(4), 931-942.

Buttle, H., & Raymond, J. E. (2003). High familiarity enhances visual change detection for face stimuli. *Perception & psychophysics*, *65*(8), 1296-1306.

Caldara, R., Schyns, P., Mayer, E., Smith, M. L., Gosselin, F., & Rossion, B. (2005). Does prosopagnosia take the eyes out of face representations? Evidence for a defect in representing diagnostic facial information following brain damage. *Journal of cognitive neuroscience*, *17*(10), 1652-1666.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67-74). IEEE.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 172-186.

Carey, S. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 95-103.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

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.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *IEEE transactions on biometrics, behavior, and identity science*, *3*(1), 101-111.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of vision*, *5*(9), 1-1.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision research*, *141*, 204-216.

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

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., ... & Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature, 547*(7663), 340-344.

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., ... & Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature*, *547*(7663), 340-344.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613-617.

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

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & psychophysics*, *58*(4), 602-612.

Crookes, K., & McKone, E. (2009). Early maturity of face recognition: No childhood development of holistic processing, novel face encoding, or face-space. *Cognition*, *111*(2), 219-247.

Crowston, K. (2012). Amazon Mechanical Turk: A research tool for organisations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches* (pp. 210-221). Springer, Berlin, Heidelberg.

Da Costa, A. L. N. T., & Do, M. N. (2014). A retina-based perceptually lossless limit and a Gaussian foveation scheme with loss control. *IEEE Journal of Selected Topics in Signal Processing*, *8*(3), 438-453.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology*, *32*(4), 331-331.

David, A. S. (1993). Spatial and selective attention in the cerebral hemispheres in depression, mania, and schizophrenia. *Brain and Cognition*, *23*(2), 166-180.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(4), 482-505.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, *30*(6), 827-840.

De Lillo, M., Foley, R., Fysh, M. C., Stimson, A., Bradford, E. E., Woodrow-Hill, C., & Ferguson, H. J. (2021). Tracking developmental differences in real-world social attention across adolescence, young adulthood and older adulthood. *Nature Human Behaviour*, 1-10.

De Renzi, E., & Di Pellegrino, G. (1998). Prosopagnosia and alexia without object agnosia. *Cortex*, *34*(3), 403-415.

DeGutis, J., Cohan, S., & Nakayama, K. (2014). Holistic face training enhances face processing in developmental prosopagnosia. *Brain*, *137*(6), 1781-1798.

DeGutis, J., Cohan, S., Kahn, D. A., Aguirre, G. K., & Nakayama, K. (2013). Facial expression training improves emotion recognition and changes neural tuning in a patient with acquired emotion recognition deficits and prosopagnosia. *Journal of Vision*, *13*(9), 993-993.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87-100.

Delaunay, B. (1934). Sur la sphere vide. Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk, 7(793-800), 1-2.

Dessimoz, D., & Champod, C. (2008). Linkages between biometrics and forensic science. In *Handbook of biometrics* (pp. 425-459). Springer, Boston, MA.

Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialisation emerges spontaneously in deep neural networks. *Sci. Adv. 8*, eabl8913. https://doi.org/10.1126/sciadv.abl8913.

Dodge, S., & Karam, L. (2016, June). Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)* (pp. 1-6). IEEE.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs outperform individuals and provide a route to training. *British journal of psychology*, *106*(3), 433-445.

Duchaine, B. & Nakayama, K. The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia 44*, 576–585 (2006).

Duchaine, B. C., Wendt, T. N. V., New, J., & Kulomäki, T. (2003). Dissociations of visual recognition in a developmental agnosic: Evidence for separate developmental processes. *Neurocase*, *9*(5), 380-389.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of cognitive neuroscience*, *17*(2), 249-261.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576-585.

Dunn, J. D., de Lima Varela, V. P., Nicholls, V. I., Papinutto, M., White, D., & Miellet, S. (2022). Face information sampling in super-recognisers. *Psychological Science*. https://doi.org/10.31234/osf.io/z2k4a

Dunn, J. D., Kemp, R. I., & White, D. (2018). Search templates that incorporate within-face variation improve visual search for faces. *Cognitive Research: Principles and Implications*, *3*(1), 1-11.

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognisers. *PloS one*, *15*(11), e0241747.

Ellis, H. D., & Young, A. W. (1988). Training in face-processing skills for a child with acquired prosopagnosia. *Developmental Neuropsychology*, *4*(4), 283-294.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception?. *Psychological review*, *105*(3), 482.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101-107.

Ferber, S., & Murray, L. J. (2005). Are perceptual judgments dissociated from motor processes?—A prism adaptation study. *Cognitive Brain Research*, *23*(2-3), 453-456.

Foulsham, T. (2020). Beyond the picture frame: the function of fixations in interactive tasks. *Psychology of learning and motivation–Advances in research and theory*, *73*, 33-58.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, 117(3), 319–331

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920-1931.

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

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society open science*, *4*(6), 170249.

Fysh, M. C., & Bindemann, M. (2018). Human–computer interaction in face matching. *Cognitive science*, *42*(5), 1714-1732.

Fysh, M. C., Trifonova, I. V., Allen, J., McCall, C., Burton, A. M., & Bindemann, M. (2021). Avatars with faces of real people: A construction method for scientific experiments in virtual reality. *Behavior Research Methods*, 1-15.

Gamer, M., & Büchel, C. (2009). Amygdala activation predicts gaze toward fearful eyes. *Journal of Neuroscience*, *29*(28), 9123-9126.

Germine, L. et al. Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Curr. Biol., 25*, 2684–2689 (2015).

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201-210.

Gibbs, J. W. (1898). Fourier's series. *Nature*, *59*(1522), 200-200.

Gibson, J. J. (1979). The ecological approach to visual perception. Houghton, Mifflin and Company.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32-41.

Goffaux, V., Hault, B., Michel, C., Vuong, Q. C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception*, *34*(1), 77-86.

Gonzalez, R. C. (2009). *Digital image processing*. Pearson education india.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, *41*(17), 2261-2271.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681.

Gregory, N. J., Bolderston, H., & Antolin, J. V. (2019). Attention to faces and gaze following in social anxiety: Preliminary evidence from a naturalistic eye-tracking investigation. *Cognition and Emotion, 33*(5), 931-942.

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., ... & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications*, *10*(1), 1-13.

Grother, P. J., Ngan, M., & Hanaoka, K., (2014). *Face recognition vendor test (frvt)*. US Department of Commerce, National Institute of Standards and Technology.

Grother, P., Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face Recognition Vendor Test (FRVT) Part 2: Identification*. US Department of Commerce, National Institute of Standards and Technology.

Growns, B., Dunn, J. D., Mattijssen, E. J., Quigley-McBride, A., & Towler, A. (2022). Match me if you can: Evidence for a domain-general visual comparison ability. *Psychonomic Bulletin & Review*, 1-16.

Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, *42*, 279-297.

Haensel, J. X., Danvers, M., Ishikawa, M., Itakura, S., Tucciarelli, R., Smith, T. J., & Senju, A. (2020). Culture modulates face scanning during dyadic social interactions. *Scientific Reports*, *10*(1), 1-11.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in cognitive sciences*, *4*(9), 330-337.

Hancock, P.J.B. et al. (1996) Face processing: human perception and principal component analysis. *Mem. Cognit., 24*, 26–40

Haoxiang, L. et al. (2015) A convolutional neural network cascade for face detection. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334, IEEE. https://doi.org/10. 1109/CVPR.2015.7299170.

Havard, C. (2007). *Eye movement strategies during face matching* (Doctoral dissertation, University of Glasgow).

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223-233.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Heller, W., & Levy, J. (1981). Perception and expression of emotion in right-handers and left-handers. *Neuropsychologia*, *19*(2), 263-272.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & cognition*, *33*(1), 98-106.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology, 15*, 445–464.

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin & Review*, *27*(5), 856-881.

Hessels, R. S., Benjamins, J. S., Cornelissen, T. H., & Hooge, I. T. (2018). A validation of automatically-generated areas-of-interest in videos of a face for eye-tracking research. *Frontiers in psychology*, *9*, 1367.

Hessels, R. S., van Doorn, A. J., Benjamins, J. S., Holleman, G. A., & Hooge, I. T. (2020). Task-related gaze control in human crowd navigation. *Attention, Perception, & Psychophysics*, *82*(5), 2482-2501.

Hill, M. Q., Parde, C. J., Castillo, C. D., Colon, Y. I., Ranjan, R., Chen, J. C., ... & O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), 522-529.

Hill, H., Roodenrys, S., & Clifford, C. (2019, April). Poor Image Quality Leads to a Conservative Bias When Matching Facial Identity. In *PERCEPTION* (Vol. 48, pp. 76-77). 1 Oliver's yard, 55 City Road, London EC1Y 1SP, England: Sage Publications Ltd.

Hong, H. et al. (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci. 19*, 613–622

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389.

Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *Plos one*, *15*(8), e0237855.

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological science*, *19*(10), 998-1006.

Hu, Y., Jackson, K., Yates, A., White, D., Phillips, P. J., & O'Toole, A. J. (2017). Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification. *Visual Cognition*, *25*(4-6), 492-506.

Huang, G.B. et al. (2012) Learning hierarchical representations for face verification with convolutional deep belief networks. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2518–2525, IEEE.

Huey, E. B. (1898). Preliminary experiments in the physiology and psychology of reading. *The American Journal of Psychology*, *9*(4), 575-586.

Iskra, A., & Tomc, H. G. (2016). Eye-tracking analysis of face observing and face recognition. *Journal of Graphic Engineering and Design*, *7*(1), 5-11.

Itz, M. L., Golle, J., Luttmann, S., Schweinberger, S. R., & Kaufmann, J. M. (2017). Dominance of texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, *108*(2), 369-396.

Itz, M. L., Schweinberger, S. R., & Kaufmann, J. M. (2018). Familiar face priming: The role of second-order configuration and individual face recognition abilities. *Perception*, *47*(2), 185-196.

Jain, A. K., & Ross, A. (2015). Bridging the gap: from biometrics to forensics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1674), 20140254.

Jain, A. K., Klare, B., & Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE multimedia*, *19*(1), 20.

Jammal, A. A., Thompson, A. C., Mariottoni, E. B., Berchuck, S. I., Urata, C. N., Estrela, T., ... & Medeiros, F. A. (2020). Human versus machine: comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *American journal of ophthalmology*, *211*, 123-131.

Jang, H., & Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of vision*, *21*(12), 6-6.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PloS one*, *10*(5), e0123780.

Jeantet, C., Caharel, S., Schwan, R., Lighezzolo-Alnot, J., & Laprevote, V. (2018). Factors influencing spatial frequency extraction in faces: A review. *Neuroscience & Biobehavioral Reviews*, *93*, 123-138.

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, *109*(4), 724-735.

Jenkins, R., & Burton, A. M. (2008). Limitations in facial identification: The evidence. Justice of the Peace, 172, 4–6.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. Philosophical Transactions of the Royal Society B, 366, 1671–1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323.

Jones, R. D., & Tranel, D. (2001). Severe developmental prosopagnosia in a child with superior intellect. *Journal of Clinical and Experimental Neuropsychology*, *23*(3), 265-273.

Jongerius, C., Callemein, T., Goedemé, T., Van Beeck, K., Romijn, J. A., Smets, E. M. A., & Hillen, M. A. (2021). Eye-tracking glasses in face-to-face interactions: Manual versus automated assessment of areas-of-interest. *Behavior Research Methods*, *53*(5), 2037-2048.

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proceedings of the National Academy of Sciences*, *119*(27), e2115047119.

Kassner, M., Patera, W., & Bulling, A. (2014, September). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 1151-1160).

Kaufmann, J. M., Schulz, C., & Schweinberger, S. R. (2013). High and low performers differ in the use of shape information for face recognition. *Neuropsychologia*, *51*(7), 1310-1319.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4873-4882).

Kemp, R., McManus, C., & Pigott, T. (1990). Sensitivity to the displacement of facial features in negative and inverted images. Perception, 19(4), 531-543.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(3), 211-222.

Kennedy, D. P., D'Onofrio, B. M., Quinn, P. D., Bölte, S., Lichtenstein, P., & Falck-Ytter, T. (2017). Genetic influence on eye movements to complex scenes at short timescales. *Current Biology*, *27*(22), 3554-3560.

Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, *19*(1), 52-56.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, *20*(3), 226-239.

Knoche, M., & Rigoll, G. (2023, July). Tackling Face Verification Edge Cases: In-Depth Analysis and Human-Machine Fusion Approach. In *2023 18th International Conference on Machine Vision and Applications (MVA)* (pp. 1-5).

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological science*, *21*(1), 38-43.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J. P., Baciu, M., Kahane, P., ... & Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, *1*(1), 1-12.

Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., & Shao, L. (2020). Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, *23*, 2086-2099.

Laidlaw, K. E., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, *108*(14), 5548-5553.

Lao, J., Miellet, S., Pernet, C., Sokhn, N., & Caldara, R. (2017). iMap4: An open source toolbox for the statistical fixation mapping of eye movement data with linear mixed modeling. *Behavior research methods*, *49*(2), 559-575.

Lê, S., Raufaste, E., & Démonet, J. F. (2003). Processing of normal, inverted, and scrambled faces in a patient with prosopagnosia: behavioural and eye tracking data. *Cognitive Brain Research*, *17*(1), 26-35.

Lê, S., Raufaste, E., Roussel, S., Puel, M., & Démonet, J. F. (2003). Implicit face perception in a patient with visual agnosia? Evidence from behavioural and eye-tracking analyses. *Neuropsychologia*, *41*(6), 702-712.

Lee, W. J., Wilkinson, C., Memon, A., & Houston, K. (2009). Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage. *AXIS, 1*, 19–28.

Li, J., Tian, M., Fang, H., Xu, M., Li, H., & Liu, J. (2010). Extraversion predicts individual differences in face recognition. *Communicative & Integrative Biology*, *3*(4), 295-298.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current biology*, *24*(21), 2569-2574.

Liu, C. H., Seetzen, H., Burton, A. M., & Chaudhuri, A. (2003). Face recognition is robust with incongruent image resolution: Relationship to security video images. *Journal of Experimental Psychology: Applied, 9*, 33–41.

Liu, X., Pedersen, M., & Wang, R. (2022). Survey of natural image enhancement techniques: Classification, evaluation, challenges, and perspectives. *Digital Signal Processing*, 103547.

Liu, Y., Zhou, L., Zhang, P., Bai, X., Gu, L., Yu, X., ... & Hancock, E. R. (2022). Where to Focus: Investigating Hierarchical Attention Relationship for Fine-Grained Visual Classification. In *European Conference on Computer Vision* (pp. 57-73). Springer, Cham.

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

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, *9*(70), 209-219.

Maaten, L. V. D., & Hinton, G. (2008). Visualising data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.

Mann, M., & Smith, M. (2017). Automated facial recognition technology: Recent developments and approaches to oversight. *UNSWLJ*, *40*, 121.

Matthews, C. M., & Mondloch, C. J. (2018). Improving identity matching of newly encountered faces: Effects of multi-image training. *Journal of Applied Research in Memory and Cognition*, *7*(2), 280-290.

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive research: principles and implications*, *3*(1), 1-15.

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., ... & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports*, *9*(1), 1-13.

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

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

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. Acta Psychologica, 137, 83–89.

Megreya, A. M., Bindemann, M., Havard, C., & Burton, A. M. (2012). Identity-lineup location influences target selection: Evidence from eye movements. *Journal of Police and Criminal Psychology*, *27*(2), 167-178.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology, 64*, 1473–1483

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision*, *14*(7), 6-6.

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

Mele, M. L., & Federici, S. (2012). Gaze and eye-tracking solutions for psychological research. *Cognitive Processing*, *13*(1), 261-265.

Meuwly, D., & Veldhuis, R. (2012, September). Forensic biometrics: From two communities to one discipline. In *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)* (pp. 1-12). IEEE.

Miellet, S., Vizioli, L., He, L., Zhou, X., & Caldara, R. (2013). Mapping face recognition information use across cultures. *Frontiers in psychology*, *4*, 34.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior research methods, instruments, & computers*, *36*(4), 630-633.

Moreton, R., Havard, C., Strathie, A., & Pike, G. (2021). An international survey of applied face-matching training courses. *Forensic Science International*, *327*, 110947.

Morrisey, M. N., Hofrichter, R., & Rutherford, M. D. (2019). Human faces capture attention and attract first saccades without longer fixation. *Visual Cognition*, *27*(2), 158-170.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577.

Nador, J. D., Zoia, M., Pachai, M. V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific reports*, *11*(1), 1-11.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision research*, *39*(23), 3824-3833.

Nasiopoulos, E., Risko, E. F., & Kingstone, A. (2015). Social attention, social presence, and the dual function of gaze. In *The many faces of social attention* (pp. 129-155). Springer, Cham.

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In *Face processing: Systems, disordersdisorders, and cultural differences* (pp. 173-201). Nova Science Publishers Inc.

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences*, 60(2), 331-340.

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *37*(5), 1149-1155.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, *22*(9), 794-809.

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 (TAP)*, *9*(4), 1-13.

O'Toole, A.J. et al. (1999) Three-dimensional shape and two dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing. *Vis. Res. 39*, 3145–3155.

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.

Palanica, A., & Itier, R. J. (2015). Eye gaze and head orientation modulate the inhibition of return for faces. *Attention, Perception, & Psychophysics*, *77*(8), 2589-2600.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia, 45*(1), 75-92.

Papinutto, M., Lao, J., Ramon, M., Caldara, R., & Miellet, S. (2017). The facespan—the perceptual span for face recognition. *Journal of vision*, *17*(5), 16-16.

Parde, C.J.et al.(2017) Face and image representation in deep CNN features. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 673–680, IEEE

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Desmaison, A. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8026-8037).

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314-E3323.

Phillips, P. J. (2017, May). A cross benchmark assessment of a deep convolutional neural network for face recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 705-710). IEEE.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & Chen, J. C. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176.

Press Association (2018, May 5). *Welsh police wrongly identify thousands as potential criminals.* The Guardian. https://www.theguardian.com/uk-news/2018/may/05/ welsh-police-wrongly-identify-thousands-as-potential-criminals

Ramon, M. (2021). Super-Recognizers–a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809.

Ramon, M., & Rossion, B. (2010). Impaired processing of relative distances between features and of the eye region in acquired prosopagnosia—two sides of the same holistic coin?. *cortex*, *46*(3), 374-389.

Ramon, M., Bobak, A. K. & White, D. Super-recognizers: from the lab to the world and back again. Br. J. Psychol. 110, 461–479 (2019).

Ramon, M., Busigny, T., & Rossion, B. (2010). Impaired holistic processing of unfamiliar individual faces in acquired prosopagnosia. *Neuropsychologia*, *48*(4), 933-944.

Ranjan, R. et al. (2007) An all-in-one convolutional neural network for face analysis. IEEE Trans. *Pattern Anal. Mach. Intell. 27*, 1–6

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, *12*(1), 1-14.

Rennels, J. L., & Davis, R. E. (2008). Facial experience during the first year. *Infant Behavior and Development*, *31*(4), 665-678.

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive psychology*, *19*(4), 473-497.

Rhodes, G., Robbins, R., Jaquet, E., McKone, E., Jeffery, L., & Clifford, C. W. (2005). Adaptation and face perception: How aftereffects implicate norm-based coding of faces. In *Fitting the mind to the world: Adaptation and after-effects in high-level vision*. Oxford University Press.

Riby, D. M., & Hancock, P. J. (2009). Do faces capture the attention of individuals with Williams syndrome or autism? Evidence from tracking eye movements. *Journal of Autism and Developmental Disorders*, *39*(3), 421-431.

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science*, *24*(11), 2235-2243.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological bulletin*, *140*(5), 1281.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological science*, *22*(4), 464-471.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of vision*, *15*(9), 15-15.

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., ... & Gauthier, I. (2019). Individual differences in object recognition. *Psychological Review*, *126*(2), 226.

Risko, E. F., Richardson, D. C. & Kingstone, A. Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science, 25*(1), 70–74 (2016).

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897-905.

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, *12*(1), 94-99.

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), e0150036.

Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). Human attention in fine-grained classification. *arXiv preprint arXiv:2111.01628*.

Rösler, L., End, A., & Gamer, M. (2017). Orienting towards social features in naturalistic scenes is reflexive. *PLoS One, 12*(7), e0182037.

Royer, J., Blais, C., Charbonneau, I., Déry, K., Tardif, J., Duchaine, B., ... & Fiset, D. (2018). Greater reliance on the eye region predicts better face recognition ability. *Cognition*, *181*, 12-20.

Royer, J., Blais, C., Gosselin, F., Duncan, J., & Fiset, D. (2015). When less is more: Impact of face processing ability on recognition of visually degraded faces. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(5), 1179.

Russell, R., Duchaine, B. & Nakayama, K. Super-recognizers: people with extraordinary face recognition ability. *Psychon. Bull. Rev., 16*, 252–257 (2009).

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252-257.

Sankaranarayanan, S. et al. (2016) Triplet probabilistic embedding for face verification and clustering. In 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS 2016), pp. 1–8, IEEE

Schroff, F. et al. (2015) FaceNet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823, IEEE

Schwarzer, G., Huber, S., Grueter, M., Grueter, T., Groß, C., Hipfel, M., & Kennerknecht, I. (2007). Gaze behaviour in hereditary prosopagnosia. *Psychological research*, *71*(5), 583-590.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological science*, *13*(5), 402-409.

Scott, H., Batten, J. P., & Kuhn, G. (2019). Why are you looking at me? It's because I'm talking, but mostly because I'm staring or not doing much. I(1), 109–118.

Seckiner, D., Mallett, X., Roux, C., Meuwly, D., & Maynard, P. (2018). Forensic image analysis–CCTV distortion and artefacts. *Forensic science international*, *285*, 77-85.

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887-12892.

Shirama, A. (2012). Stare in the crowd: Frontal face guides overt attention independently of its gaze direction. *Perception*, *41*(4), 447-459.

Simonyan, K. et al. (2013) Fisher vector faces in the wild. In Procedings of the British Machine Vision Conference 2013, BMVA. Vol. 2, pp. 8.1–8.12, https://doi.org/10.5244/C.27.8.

Singer N. (2018, Jul 26). *Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers.* The New York Times. https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html

Slessor, G., Riby, D. M., & Finnerty, A. N. (2013). Age-related differences in processing face configuration: The importance of the eye region. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(2), 228-231.

Solman, G. J., Foulsham, T., & Kingstone, A. (2017). Eye and head movements are complementary in visual selection. *Royal Society Open Science*, *4*(1), 160569.

Stephan, B. C. M., & Caine, D. (2009). Aberrant pattern of scanning in prosopagnosia reflects impaired face processing. *Brain and Cognition*, *69*(2), 262-268.

Sugden, N. A., & Moulson, M. C. (2017). Hey baby, what's "up"? One-and 3-month-olds experience faces primarily upright but non-upright faces offer the best views. *Quarterly Journal of Experimental Psychology*, *70*(5), 959-969.

Sugden, N. A., & Moulson, M. C. (2019). These are the people in your neighbourhood: Consistency and persistence in infants' exposure to caregivers', relatives', and strangers' faces across contexts. *Vision Research*, *157*, 230-241.

Sun, Y. et al. (2014) Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1891–1898, IEEE.

Sunday, M. A., Richler, J. J., & Gauthier, I. (2017). Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Attention, Perception, & Psychophysics*, *79*(5), 1453-1465.

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. *Current opinion in neurobiology*, *23*(3), 423-429.

Sutherland, C. A., Burton, N. S., Wilmer, J. B., Blokland, G. A., Germine, L., Palermo, R., ... & Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by

environments, not genes. *Proceedings of the National Academy of Sciences*, *117*(19), 10218-10224.

Taigman, Y. et al. (2014) DeepFace: closing the gap to human level performance in face verification. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708, IEEE. https://doi.org/10.1109/CVPR.2014.220

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly journal of experimental psychology*, *46*(2), 225-245.

Tardif, J., Morin Duchesne, X., Cohan, S., Royer, J., Blais, C., Fiset, D., Duchaine, B., & Gosselin, F. (2019). Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychological science*, *30*(2), 300-308.

The jamovi project. (2021). jamovi. Retrieved from https://www.jamovi.org
Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657-665.

Thomaz, C. E., Amaral, V., Giraldi, G. A., Gillies, D. F., & Rueckert, D. (2017). Is human face processing a feature-or pattern-based task? evidence using a unified computational method driven by eye movements. *arXiv preprint arXiv:1709.01182*.

Thomaz, C. E., Boardman, J. P., Counsell, S., Hill, D. L., Hajnal, J. V., Edwards, A. D., ... & Rueckert, D. (2007). A multivariate statistical analysis of the developing human brain in preterm infants. *Image and Vision Computing*, *25*(6), 981-994.

Thomaz, C. E., Kitani, E. C., and Gillies, D. F. (2006). A maximum uncertainty lda based approach for limited sample size problems-with application to face recognition. Journal of the Brazilian Computer Society, 12(2):7–18.

Towler, A., Dunn, J. D., Martinez, S., Moreton, R., Eklöf, F., Ruifrok, A., Kemp, R. I., & White, D. (2021). Diverse routes to expertise in facial recognition. Pre-print: 10.31234/osf.io/fmznh

Towler, A., Kemp, R. I. & White, D. in Forensic face matching: Research and practice (ed M. Bindemann) (Oxford University Press, 2021).

Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar face matching systems in applied settings. *Face processing: systems, disorders and cultural differences. New York: Nova Science Publishing, Inc*.

Towler, A., Kemp, R. I., & White, D. (2021). Can Face Identification Ability Be Trained?. *Forensic face matching: Research and practice*, 89.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work?. *PloS one*, *14*(2), e0211037.

Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (2021). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47.

Towler, J., Fisher, K., & Eimer, M. (2017). The cognitive and neural basis of developmental prosopagnosia. *Quarterly Journal of Experimental Psychology*, *70*(2), 316-344.

Tsantani, M., & Cook, R. (2020). Normal recognition of famous voices in developmental prosopagnosia. *Scientific reports*, *10*(1), 1-11.

Tsantani, M., Gray, K. L., & Cook, R. (2020). Holistic processing of facial identity in developmental prosopagnosia. *Cortex*, *130*, 318-326.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA encode distinct types of face identity information. *Journal of Neuroscience*, *41*(9), 1952-1969.

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

Varela, V. P., Ribeiro, E., Orona, P. A., & Thomaz, C. E. (2018, October). Eye movements and human face perception: An holistic analysis and proficiency classification based on frontal 2D face images. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional* (pp. 48-57). SBC.

Vera-Rodriguez, R., Blazquez, M., Morales, A., Gonzalez-Sosa, E., Neves, J. C., & Proença, H. (2019). FaceGenderID: Exploiting gender information in DCNNs face recognition systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision research*, *141*, 217-227.

Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). ˜ Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. Journal of Vision, 12(13), 1–14.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, *115*(44), 11333-11338.

Vuilleumier, P. (2000). Faces call for attention: evidence from patients with visual extinction. *Neuropsychologia*, *38*(5), 693-700.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological science*, *23*(2), 169-177.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600-612.

White, D., & Burton, A. M. (2022). Individual differences and the multi-dimensional nature of face perception. *Nature Reviews Psychology*, 1-14.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied cognitive psychology*, *27*(6), 769-777.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PloS one*, *10*(10), e0139827.

White, D., Guilbert, D., Varela, V. P., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. *Behavior research methods*, *54*(1), 252-260.

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic bulletin & review*, *21*(1), 100-106.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one*, *9*(8), e103510.

White, D., Towler, A. & Kemp, R. I. in Forensic face matching: Research and practice (ed M. Bindemann) (Oxford University Press, 2021).

White, D., Towler, A., & Kemp, R. (2021). Understanding professional expertise in unfamiliar face matching. *Forensic face matching: Research and practice*, 62-88.

White, D., 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,* 104966.

Wilcockson, T. D., Burns, E. J., Xia, B., Tree, J., & Crawford, T. J. (2020). Atypically heterogeneous vertical first fixations to faces in a case series of people with developmental prosopagnosia. *Visual Cognition*, *28*(4), 311-323.

Williams, R. (2020, June 24). *Opinion: I was wrongfully arrested because of facial recognition. Why are police allowed to use it?.* The Washington Post. https://www.washingtonpost.com/opinions/2020/06/24/i-was-wrongfully-arrested-because-facial-recognition-why-are-police-allowed-use-this-technology/

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive neuropsychology*, *29*(5-6), 360-392.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of sciences*, *107*(11), 5238-5241.

Wolff, W. (1933). The experimental study of forms of expression. Character & Personality; A Quarterly for Psychodiagnostic & Allied Studies.

Wolley, S. (2021, May 21). *Rape with homicide complaint against eleven men over death of flight attendant Christine Dacera dismissed.* 7news. https://7news.com.au/news/world/rape-with-homicide-complaint-against-eleven-men-over-death-of-flight-attendant-christine-dacera-dismissed-c-2899190

Yan, X., Young, A. W., & Andrews, T. J. (2017). The automaticity of face perception is influenced by familiarity. Attention, Perception, & Psychophysics, 79(7), 2202-2211.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In Eye movements and vision (pp. 171-211). Springer, Boston, MA.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Young, A. W., & Burton, A. M. (2018). Are we face experts?. *Trends in Cognitive Sciences, 22(*2), 100-110.

Young, A. W., Hellawell, D. J., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16*(6), 747-759. doi:10.1068/p160747

Yovel, G., & O'Toole, A. J. (2016). Recognizing people in motion. *Trends in cognitive sciences*, *20*(5), 383-395.

Zangeneh, E., Rahmati, M., & Mohsenzadeh, Y. (2020). Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications*, *139*, 112854.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499-1503.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, *35*(4), 399-458.

Zhu et al. Heritability of the Specific Cognitive Ability of Face Perception. *Current Biology,* 2010; DOI: 10.1016/j.cub.2009.11.067

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., ... & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*(2), 137-142.

APPENDIX A

Method for pilot study to choose frequency bands in Chapter 2

Figure A.1 shows how DCNN1 (see Table 1.1 of main text) could identify the matching C2 for each C1 image as the most similar identity in the dataset. Figure A.1 shows that using FFT filtering, the matching performance of the DCNN was perfect for images containing the information below 13 cycles/face or more.

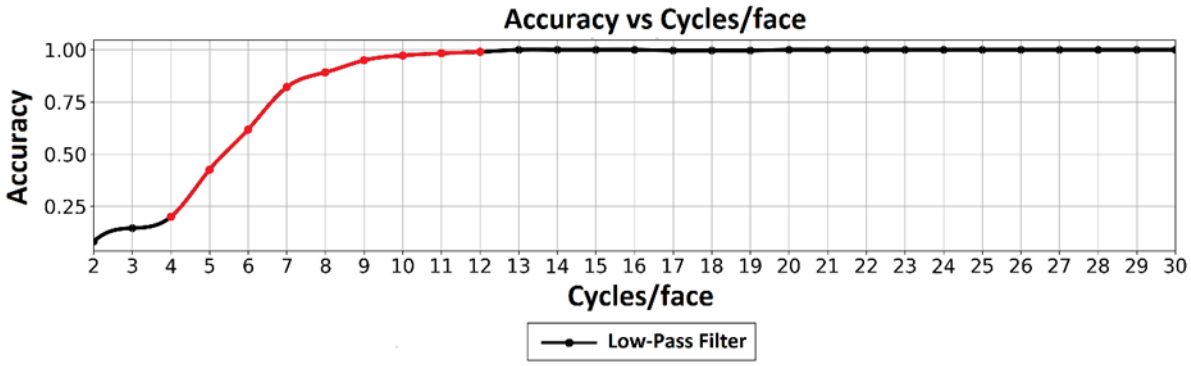


Figure A.1. Accuracy (% correct) of DCNN 1 finding match pairs in 29 manipulated versions of the database containing filtered images using high or low-pass filters on different cutoff frequencies. We used areas marked in red to create the stimuli.

We created 29 different versions of the image set using a low-pass filter, each with a different cut off frequency between 2 and 30 cycles/face, and observed how accurate DCNN1 was for all versions. We used the FFT as an ideal low-pass filter to extract specific spatial frequencies from each image.

Descriptive Table

In Experiment 1, we collected data from human participants (**Human**) and the average DCNN (**AverageAlgorithm**) in two **Experiments** (Same-Resolution and Different-Resolution) using two different **Filters** (FFT and Gaussian) and two other metrics of analysis, Yoked and Non-Yoked. Table A.1 shows the data we collected during the experiment.

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
N	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	168
				Yoked	168
			Same-Resolution	Non-Yoked	162
				Yoked	162
		Human	Different-Resolution	Non-Yoked	168
				Yoked	168
			Same-Resolution	Non-Yoked	162
				Yoked	162
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	100
				Yoked	100
			Same-Resolution	Non-Yoked	100
				Yoked	100
		Human	Different-Resolution	Non-Yoked	100
				Yoked	100
			Same-Resolution	Non-Yoked	100
				Yoked	100
Missing	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0
				Yoked	0
			Same-Resolution	Non-Yoked	0
				Yoked	0
		Human	Different-Resolution	Non-Yoked	0
				Yoked	0
			Same-Resolution	Non-Yoked	0
				Yoked	0
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	0
				Yoked	0
			Same-Resolution	Non-Yoked	0

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
				Yoked	0
		Human	Different-Resolution	Non-Yoked	0
				Yoked	0
			Same-Resolution	Non-Yoked	0
				Yoked	0
Mean	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0.591
				Yoked	0.660
			Same-Resolution	Non-Yoked	0.758
				Yoked	0.826
		Human	Different-Resolution	Non-Yoked	0.747
				Yoked	0.717
			Same-Resolution	Non-Yoked	0.812
				Yoked	0.780
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	0.687
				Yoked	0.736
			Same-Resolution	Non-Yoked	0.783
				Yoked	0.848
		Human	Different-Resolution	Non-Yoked	0.879
				Yoked	0.863
			Same-Resolution	Non-Yoked	0.888
				Yoked	0.881
Median	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0.594
				Yoked	0.663
			Same-Resolution	Non-Yoked	0.764
				Yoked	0.826
		Human	Different-Resolution	Non-Yoked	0.759
				Yoked	0.725

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
			Same-Resolution	Non-Yoked	0.832
				Yoked	0.791
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	0.688
				Yoked	0.734
			Same-Resolution	Non-Yoked	0.787
				Yoked	0.851
		Human	Different-Resolution	Non-Yoked	0.893
				Yoked	0.881
			Same-Resolution	Non-Yoked	0.908
				Yoked	0.900
Standard deviation	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0.0588
				Yoked	0.0347
			Same-Resolution	Non-Yoked	0.0559
				Yoked	0.0322
		Human	Different-Resolution	Non-Yoked	0.0866
				Yoked	0.0904
			Same-Resolution	Non-Yoked	0.0870
				Yoked	0.0916
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	0.0494
				Yoked	0.0286
			Same-Resolution	Non-Yoked	0.0479
				Yoked	0.0301
		Human	Different-Resolution	Non-Yoked	0.0751
				Yoked	0.0835
			Same-Resolution	Non-Yoked	0.0855
				Yoked	0.0896
Minimum	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0.434

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
Maximum	FFT	Human	Same-Resolution	Yoked	0.551
				Non-Yoked	0.548
			Different-Resolution	Yoked	0.710
				Non-Yoked	0.440
		AverageAlgorithm	Same-Resolution	Yoked	0.451
				Non-Yoked	0.458
			Different-Resolution	Yoked	0.462
				Non-Yoked	0.579
	Gaussian	Human	Same-Resolution	Yoked	0.663
				Non-Yoked	0.636
			Different-Resolution	Yoked	0.774
				Non-Yoked	0.620
		AverageAlgorithm	Same-Resolution	Yoked	0.557
				Non-Yoked	0.616
			Different-Resolution	Yoked	0.590
				Non-Yoked	0.728
		Human	Same-Resolution	Yoked	0.760
				Non-Yoked	0.871
			Different-Resolution	Yoked	0.897
				Non-Yoked	0.968
		AverageAlgorithm	Same-Resolution	Yoked	0.974
				Non-Yoked	0.983
			Different-Resolution	Yoked	0.959
				Non-Yoked	0.794
	Gaussian	Human	Same-Resolution	Yoked	0.799
				Non-Yoked	0.869
		AverageAlgorithm	Same-Resolution	Yoked	0.905

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
		Human	Different-Resolution	Non-Yoked	0.987
				Yoked	0.982
			Same-Resolution	Non-Yoked	1.00
				Yoked	1.00
Skewness	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	-0.0417
				Yoked	-0.293
			Same-Resolution	Non-Yoked	-0.774
				Yoked	-0.343
		Human	Different-Resolution	Non-Yoked	-0.257
				Yoked	-0.236
			Same-Resolution	Non-Yoked	-1.54
				Yoked	-1.17
	Gaussian	AverageAlgorithm	Different-Resolution	Non-Yoked	-0.143
				Yoked	-0.115
			Same-Resolution	Non-Yoked	-0.762
				Yoked	-0.320
		Human	Different-Resolution	Non-Yoked	-1.18
				Yoked	-1.26
			Same-Resolution	Non-Yoked	-1.23
				Yoked	-1.23
Std. error skewness	FFT	AverageAlgorithm	Different-Resolution	Non-Yoked	0.187
				Yoked	0.187
			Same-Resolution	Non-Yoked	0.191
				Yoked	0.191
		Human	Different-Resolution	Non-Yoked	0.187
				Yoked	0.187
			Same-Resolution	Non-Yoked	0.191

Descriptives

	Filter	Human/Algorithm	Experiment	Yoked/NonYoked	AverageAUC
	Gaussian	AverageAlgorithm	Different-Resolution	Yoked	0.191
				Non-Yoked	0.241
			Same-Resolution	Yoked	0.241
				Non-Yoked	0.241
		Human	Different-Resolution	Yoked	0.241
				Non-Yoked	0.241
			Same-Resolution	Yoked	0.241
				Non-Yoked	0.241

Table A.1 – Description of data found in Experiment 1.

ANOVA Table

Table A.2 shows the FFT version three-way ANOVA to observe the impact of AUC on three different factors: Resolution condition (Same and Different-Resolution), Participant Type (human and DCNNs), and Metric (Yoked and Non-Yoked). Table A.3 shows the same analysis for the Gaussian version of the experiment. See Figure A.2.

ANOVA – AUC - FFT VERSION

	Sum of Squares	df	Mean Square	F	p	η^2p
Resolution	4.384	1	4.38364	867.01747	< .001	0.398
ParticipantType	1.013	1	1.01321	200.39828	< .001	0.133
Metric	0.115	1	0.11492	22.72985	< .001	0.017
Resolution ✻ ParticipantType	0.859	1	0.85879	169.85486	< .001	0.115
Resolution ✻ Metric	1.37e-4	1	1.37e-4	0.02711	0.869	0.000
ParticipantType ✻ Metric	0.812	1	0.81152	160.50702	< .001	0.109
Resolution ✻ ParticipantType ✻ Metric	5.04e-5	1	5.04e-5	0.00997	0.920	0.000
Residuals	6.633	1312	0.00506			

Table A.2. FFT version ***three-way ANOVA used in Experiment 1. This analysis is to observe the impact of AUC on three different factors: Experimental condition (Same and Different-Resolution), Participant type (Human and DCNNs), and the metric (Yoked and Non-Yoked).***

ANOVA – AUC – Gaussian VERSION

	Sum of Squares	df	Mean Square	F	p	η^2p
Resolution	0.69008	1	0.69008	160.472	< .001	0.168
ParticipantType	2.61591	1	2.61591	608.311	< .001	0.434
Metric	0.10446	1	0.10446	24.291	< .001	0.030
Resolution ✻ ParticipantType	0.40688	1	0.40688	94.616	< .001	0.107
Resolution ✻ Metric	0.00852	1	0.00852	1.981	0.160	0.002
ParticipantType ✻ Metric	0.23159	1	0.23159	53.855	< .001	0.064
Resolution ✻ ParticipantType ✻ Metric	8.21e-4	1	8.21e-4	0.191	0.662	0.000
Residuals	3.40583	792	0.00430			

Table A.3. Gaussian version ***three-way ANOVA used in Experiment 1. This analysis is to observe the impact of AUC on three different factors: Experimental condition (Same and Different-Resolution), Participant type (Human and DCNNs), and the metric (Yoked and Non-Yoked).***

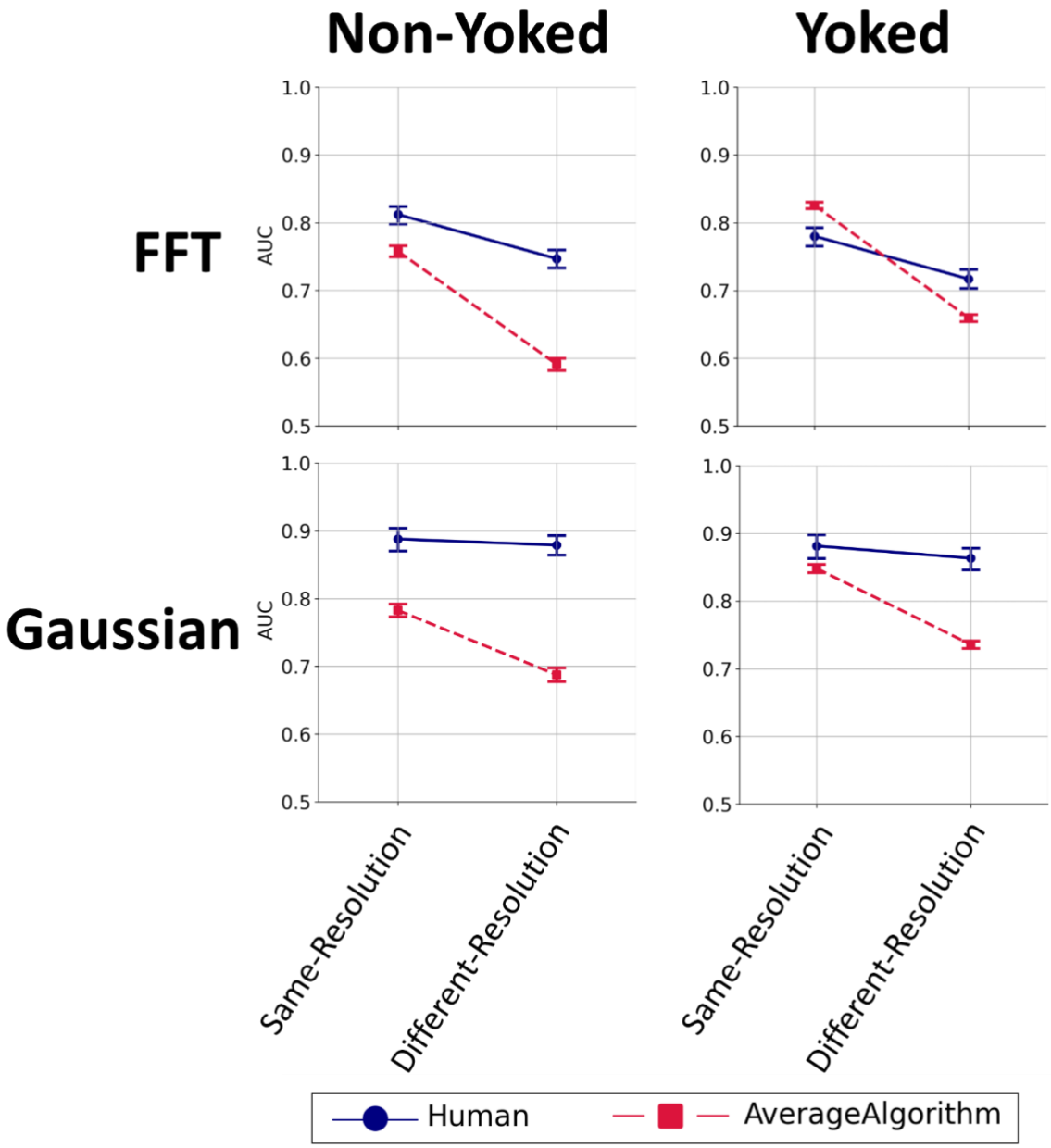


Figure A.2. Graph showing the Resolution condition between Participant Type considering the two different Metrics to compute AUC. This graph shows the FFT and Gaussian versions of the stimuli. The error bars show the 95% confidence interval.

To exclude the effect of metrics, we averaged the Yoked and Non-Yoked AUC results within participants to evaluate the impact of experimental conditions (Same and Different-Resolution) between participants in each filtering condition. See Figure 2.9. For that, we performed a two-way ANOVA to observe the effect of accuracy (average AUC) between two factors: Experimental condition (Same and Different-Resolution) and Participant type (Human and DCNNs). Table A.4 and A.5 show the ANOVA table for the FFT and Gaussian filter versions of Experiment 1. See Figure A.3.

ANOVA - AUC_AVERAGE – FFT VERSION

	Sum of Squares	df	Mean Square	F	p	η^2p
Resolution	2.192	1	2.19182	773	< .001	0.541
ParticipantType	0.507	1	0.50661	179	< .001	0.214
Resolution ✻ ParticipantType	0.429	1	0.42939	151	< .001	0.188
Residuals	*1.860*	*656*	*0.00283*			

Table A.4. ANOVA table showing the effect of accuracy (average AUC) between two factors: Experimental condition (Same and Different-Resolution) and Participant type (Human and DCNNs) in the FFT version of Experiment 1.

ANOVA - AUC_AVERAGE – Gaussian VERSION

	Sum of Squares	df	Mean Square	F	p
Resolution	0.345	1	0.34504	155.2	< .001
ParticipantType	1.308	1	1.30796	588.2	< .001
Resolution ✻ ParticipantType	0.203	1	0.20344	91.5	< .001
Residuals	0.881	396	0.00222		

Table A.5. ANOVA table showing the effect of accuracy (average AUC) between two factors: Experimental condition (Same and Different-Resolution) and Participant type (Human and DCNNs) in the Gaussian version of Experiment 1.

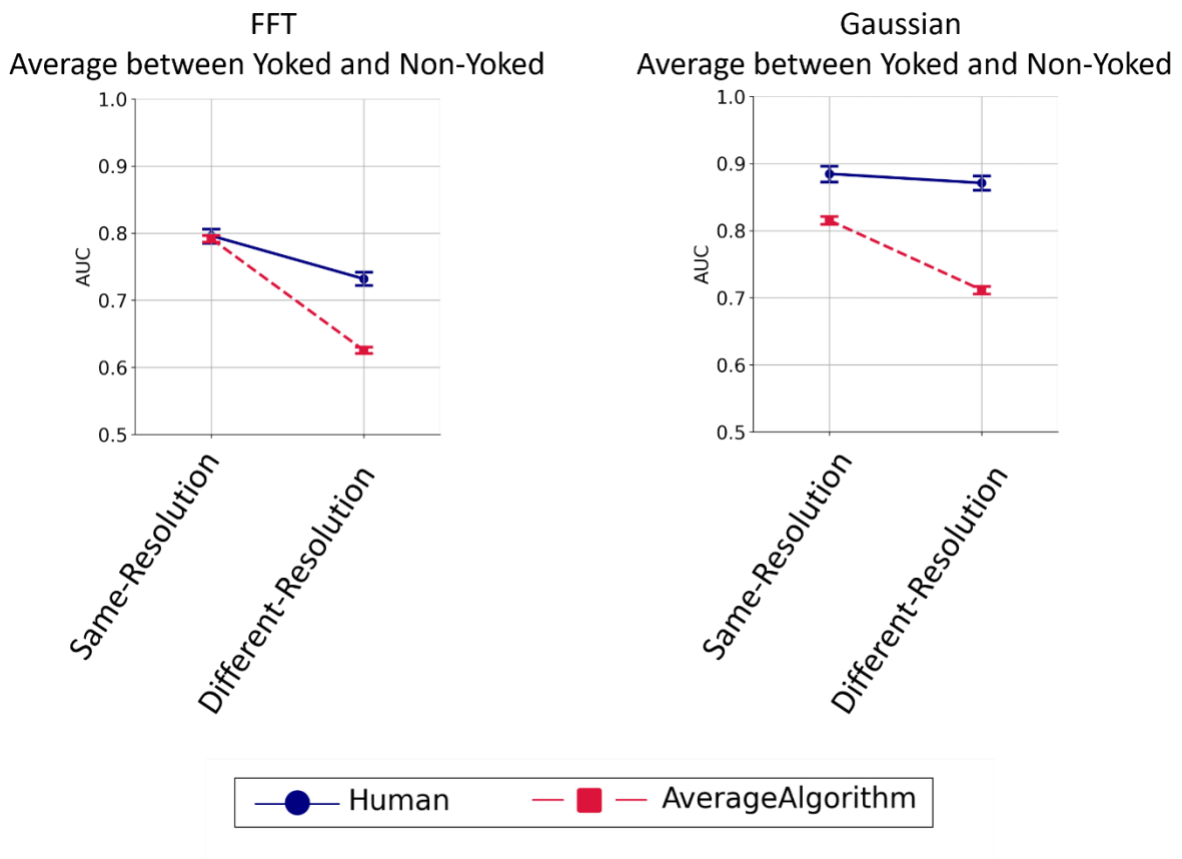


Figure A.3. Graphs show the effect of Resolution conditions for the FTT and Gaussian study versions. Here, we calculated AUC as the average result between the Non-Yoked and Yoked metrics.

Similarity Scores Distributions

We evaluated the similarity scores between humans and DCNNs performing the Same and Different-Resolution conditions using two filters, FFT and Gaussian (see the main manuscript). Figure A.4 shows the evaluation of humans and DCNN1 performing the FFT version of the experiment, while Figure A.5 shows humans and DCNN1 performing the Gaussian version of the experiment. We did not show the results containing the other DCNNs because they show similar results.

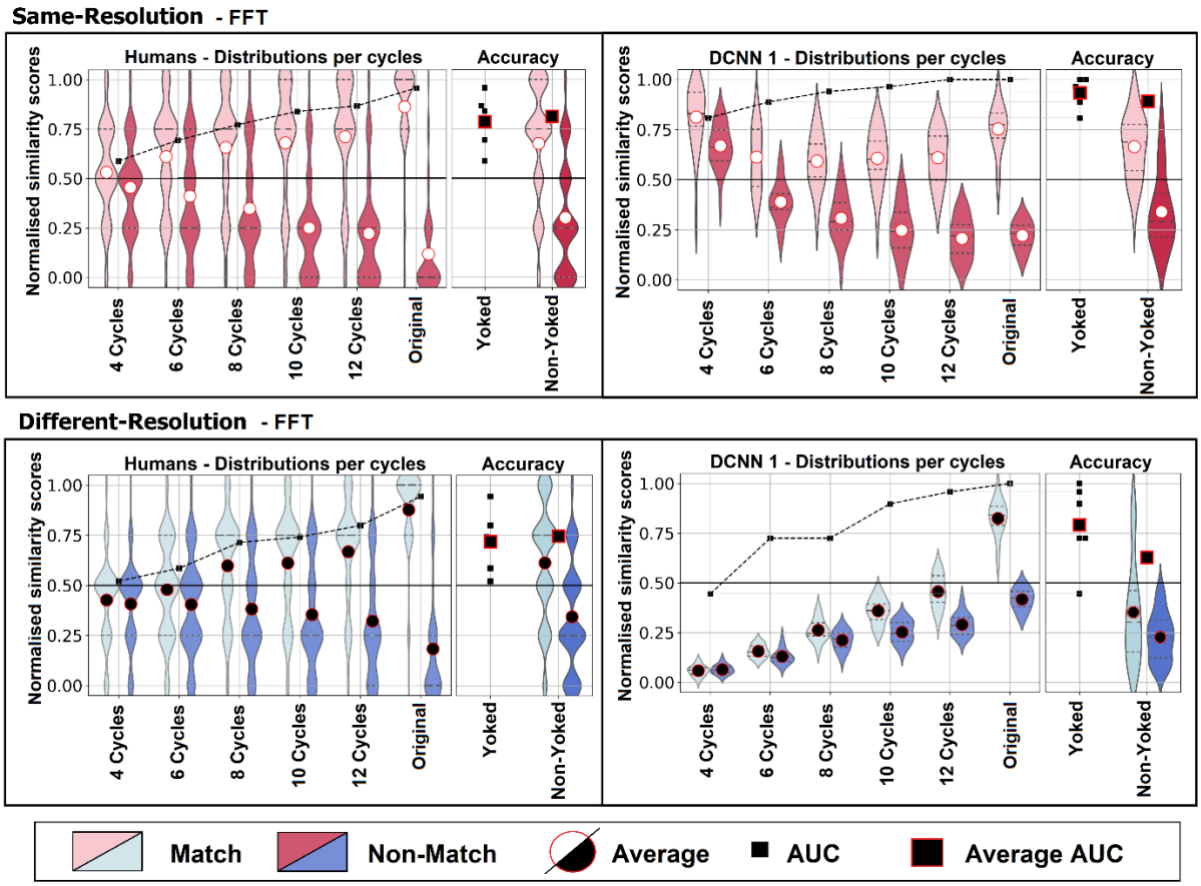


Figure A.4. Distribution of Similarity-Scores across all participants and DCNN 1 in the two experimental conditions for the FFT version. The distributions show match and non-match decisions for all spatial frequencies and the Yoked and Non-Yoked metrics results. On the left, human participants tend to respond 0.5 - or 'Do not know'- when trials are too filtered (e.g. 4 cycles/face) in both experimental conditions. In contrast, we do not observe the same effect for DCNNs (on the right side of the panel).

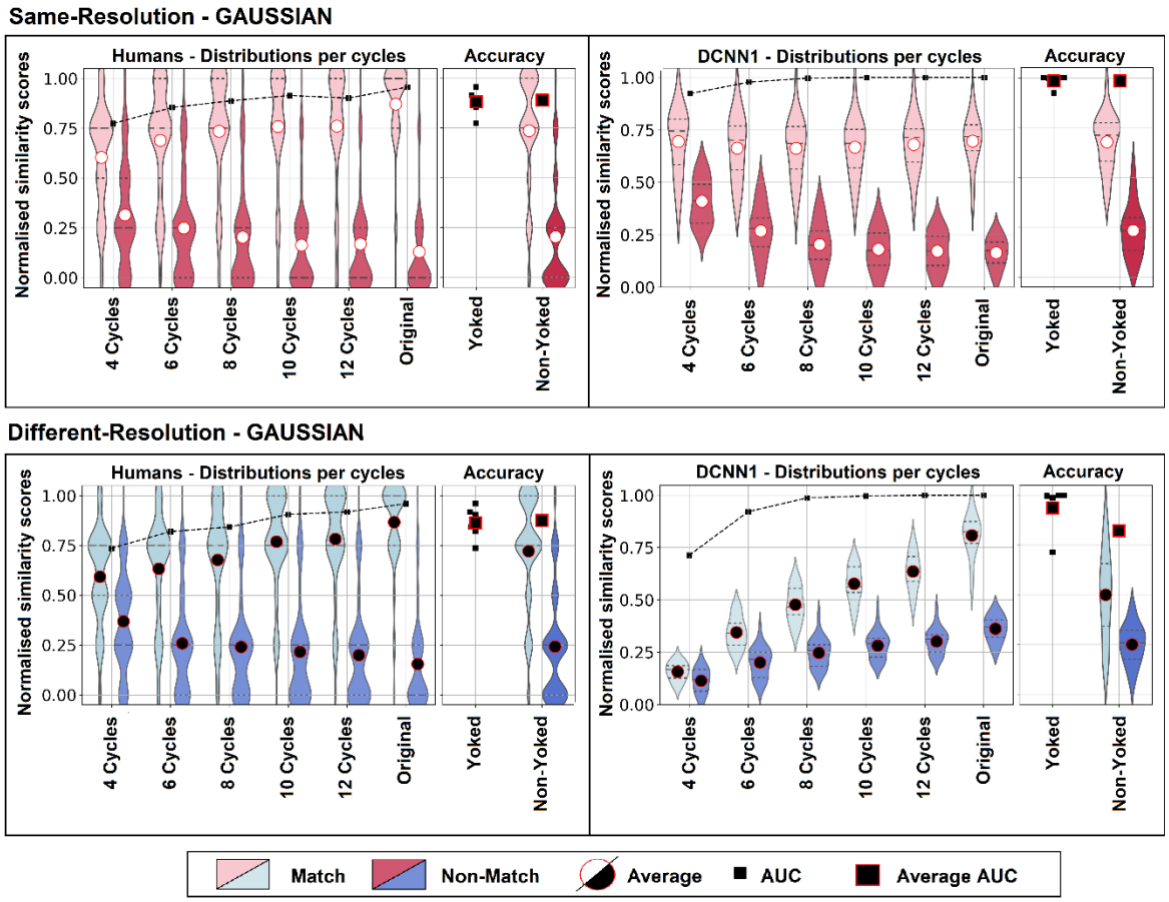


Figure A.5. Distribution of Similarity-Scores across all participants and DCNN 1 in the two experimental conditions for the Gaussian version. The distributions show match and non-match decisions for all spatial frequencies and the Yoked and Non-Yoked metrics results. On the left, human participants tend to respond 0.5 - or 'Do not know'- when trials are too filtered (e.g. 4 cycles/face) in both experimental conditions. In contrast, we do not observe the same effect for DCNNs (on the right side of the panel).

t-SNE visualisation

Figures A.6 and A.7 show the t-SNE visualisation for DCNN1 (see Table 1.1 in the main manuscript) for the manipulations using the FFT filtering and Gaussian filter, respectively, using all database images for Experiment 1.

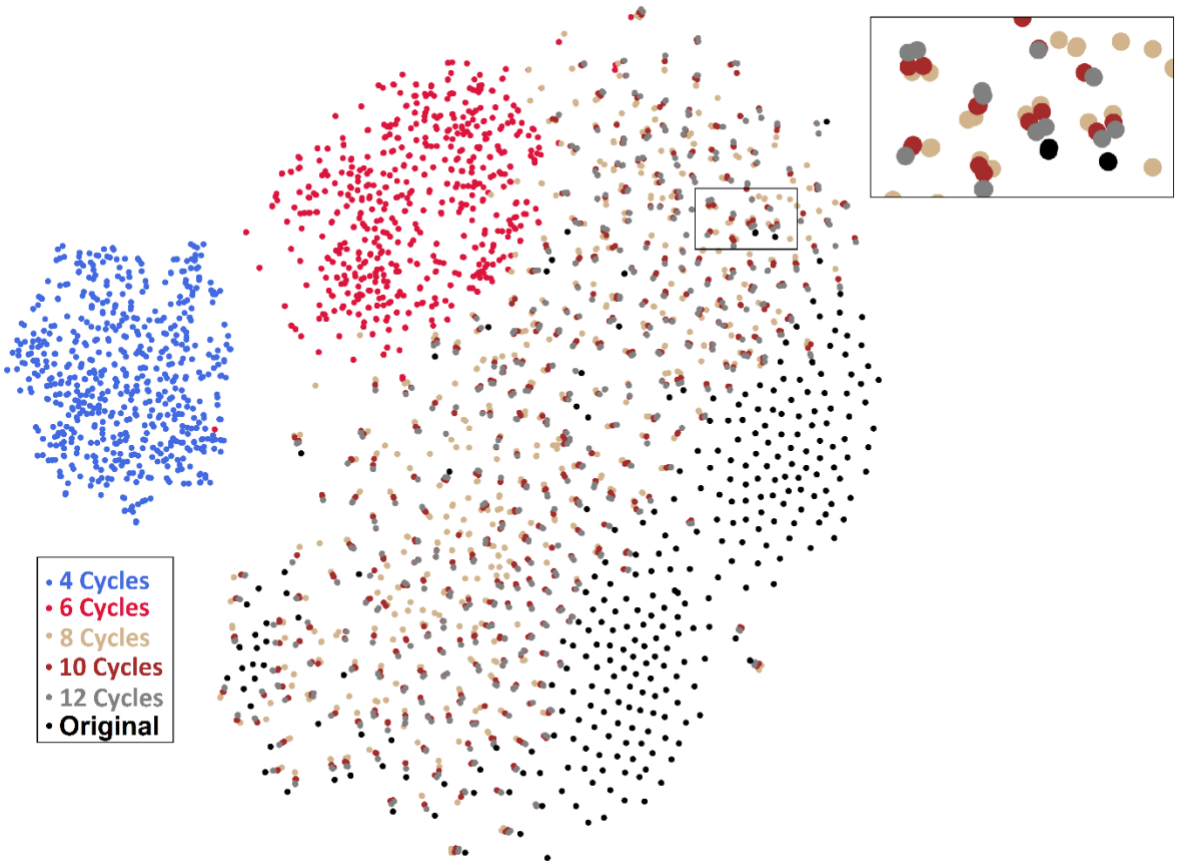


Figure A.6. Visualisation of t-SNE on DCNN 1 top-layer (2048 features) using every face present in the database used to create the stimuli (592 faces, 296 identities) in each of the six filtering conditions (4, 6, 8, 10, 12 cycles/face, and Original) using FFT filtering. It is possible to observe that black dots(Original) are – mostly - detached from the rest, suggesting the loss of identity information due to filtering. Same-Resolution condition compared distances between some dots of the same colour, while Different-Resolution compared black dots with any other colour.

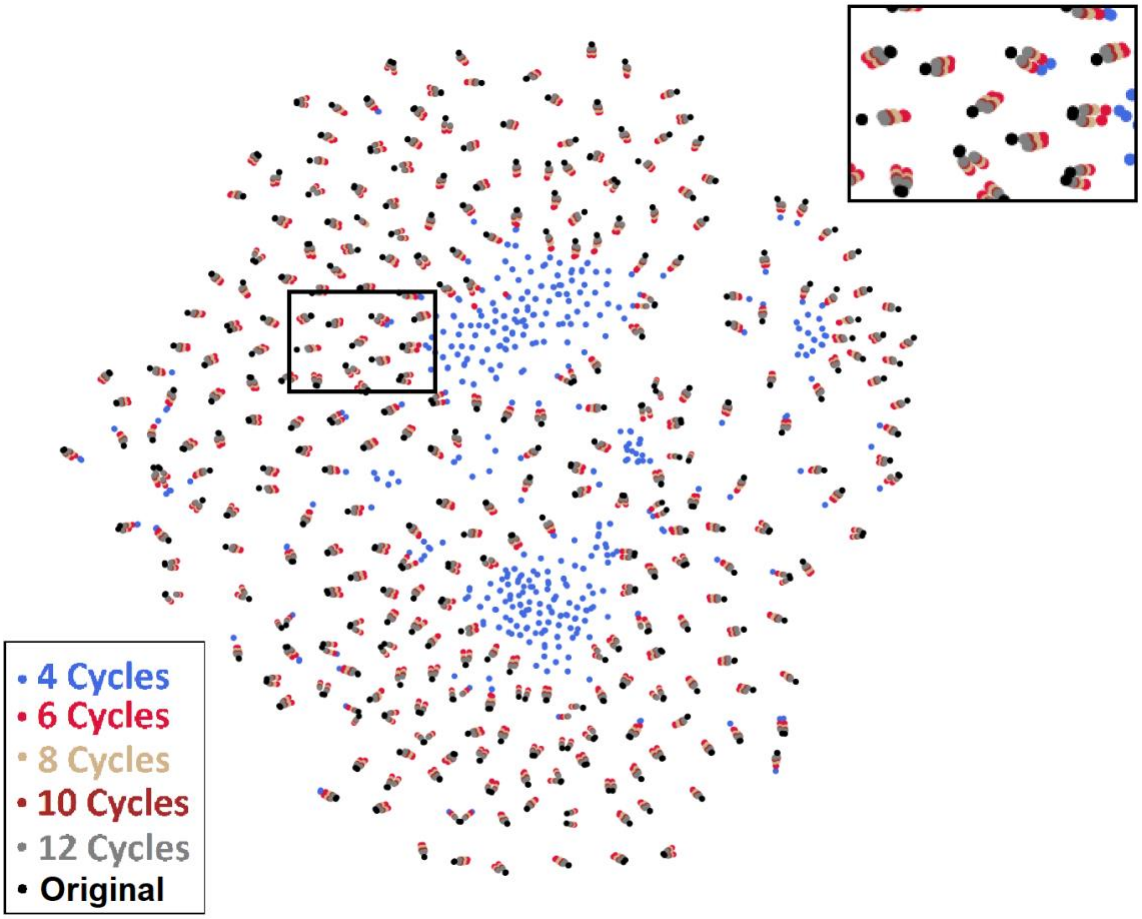


Figure A.7. Visualisation of t-SNE on DCNN 1 top-layer (2048 features) using every face present in the database used to create the stimuli (592 faces, 296 identities) in each of the six filtering conditions (4, 6, 8, 10, 12 cycles/face, and Original) using Gaussian Blur filtering. It is possible to observe that black dots (Original) remained attached to the rest, and only some highly filtered images (i.e. 4 cycles/face) lost identity information due to filtering. Same-Resolution condition compared distances between some dots of the same colour, while Different-Resolution compared black dots with any other colour.

APPENDIX B

Results for Gaussian Blur

We replicated the analysis of the main text for the Gaussian version of Experiment 2 (see main text). Figure B.1 shows the correlation between human accuracy and the fusion improvements provoked in the average DCNN performance in the two stimuli conditions. We divided Figure B.1 into two panels. The top panel shows how the decisions made by humans and the average DCNN interacted to boost performance in the Same-Resolution condition in the two fusion methods (left: Direct fusion; right: Quality Sensitive fusion). In this panel, both the Direct [rho(98)= 0.55, p< 0.001] and Quality Sensitive [rho(98)= 0.72, p< 0.001] fusion methods showed significant correlations with human performance. The bottom panel replicates the previous analysis but in the Different-Resolution condition, which shows similar correlations compared to the Same-Resolution. In this panel, both the Direct [rho(98)= 0.56, p< 0.001] and Quality Sensitive [rho(98)= 0.69, p< 0.001] fusion methods showed significant correlations with human performance. We found similar results for all DCNNs in this study.

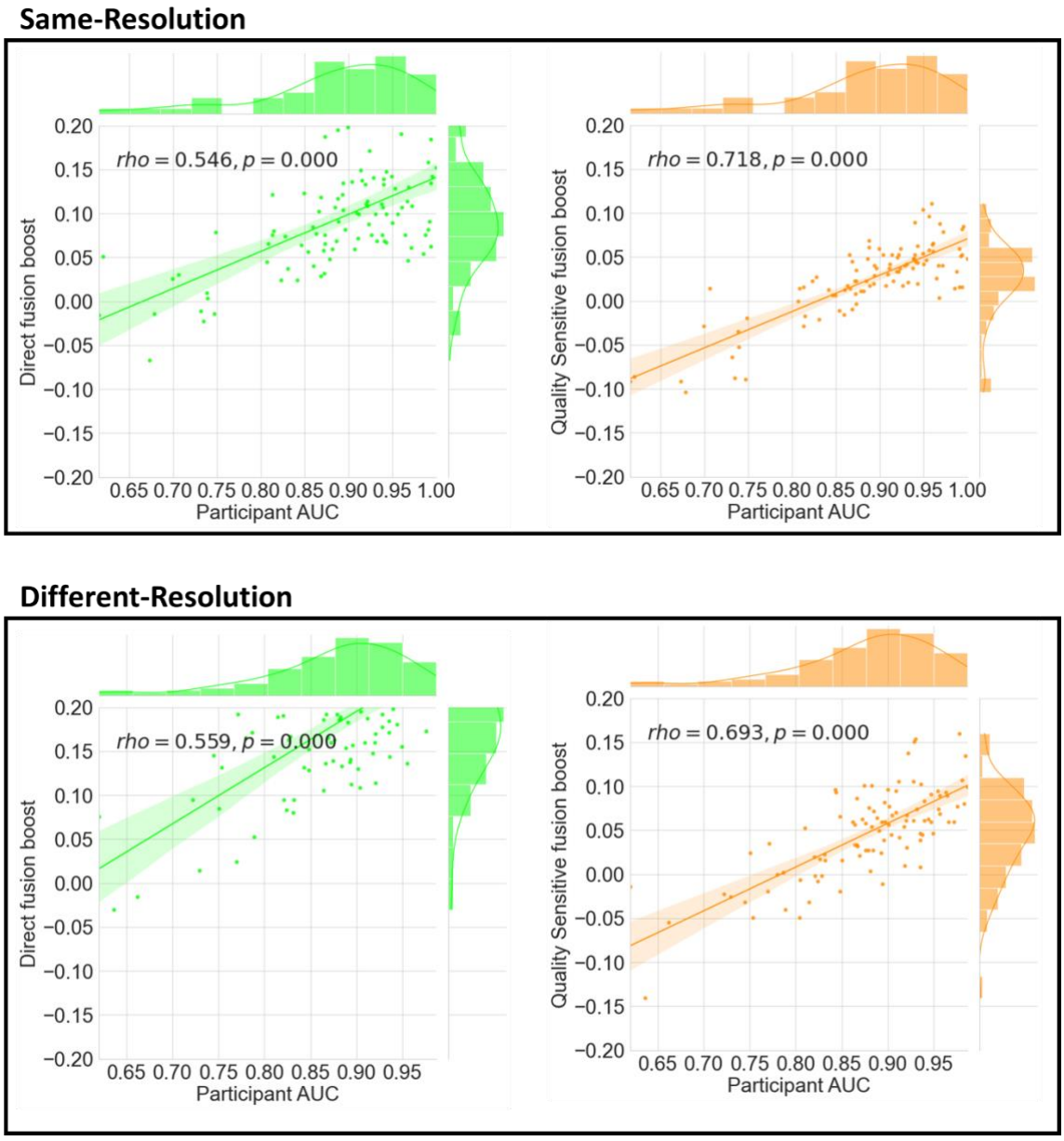


Figure B.1. Correlation between human performance (AUC) with the boost in performance in the average DCNN using the two fusion metrics. At the top panel, we show the results for the Same-Resolution, and at the bottom, the Different-Resolution condition. We calculate the boost in performance by subtracting the resulting Fusion AUC from the DCNN AUC.

We aim to understand how close human performance should be compared to DCNNs' to improve fusion decisions. Therefore, in Figure B.2, we correlate the differences in AUC found between humans and the average DCNN against the boost in performance caused by the fusion. We divided Figure B.2 into two panels. The top panel shows how the accuracy difference between humans and the average DCNN interacted to boost performance in the Same-Resolution condition in the two fusion methods (left: Direct fusion; right: Quality Sensitive fusion). In this panel, both the Direct [rho(98)= 0.92, p<

0.001] and Quality Sensitive [rho(98)= 0.88, p< 0.001] fusion methods showed significant correlations with human performance. The bottom panel replicates the previous analysis but in the Different-Resolution condition, which shows similar correlations compared to the Same-Resolution. In this panel, both the Direct [rho(167)= 0.96, p< 0.001] and Quality Sensitive [rho(98)= 0.95, p< 0.001] fusion methods showed significant correlations with human and DCNN performance. We found similar results for all DCNNs in this study. Therefore, this highly significant analysis illustrates a linear relationship between the accuracy of humans and DCNNs. Moreover, Figure B.2 shows that fusion results are better when the accuracy of humans and algorithms are -at least – similar.

This highly significant correlation shown in Figure B.2 extends our previous finding that high-performance humans might boost DCNN decisions. Interestingly, the linear trend-line that best fitted our fusion analyses in Figure B.2 predicts that the humans' entry accuracy to start increasing the DCNN's performance laid around -8% of the DCNN performance. This starting position was, considering all DCNNs, -9.5% for the Same-Resolution and -7% for the Different-Resolution conditions. See Figure B.3 for the predicted position (i.e. AUC= 0) for all DCNNs.

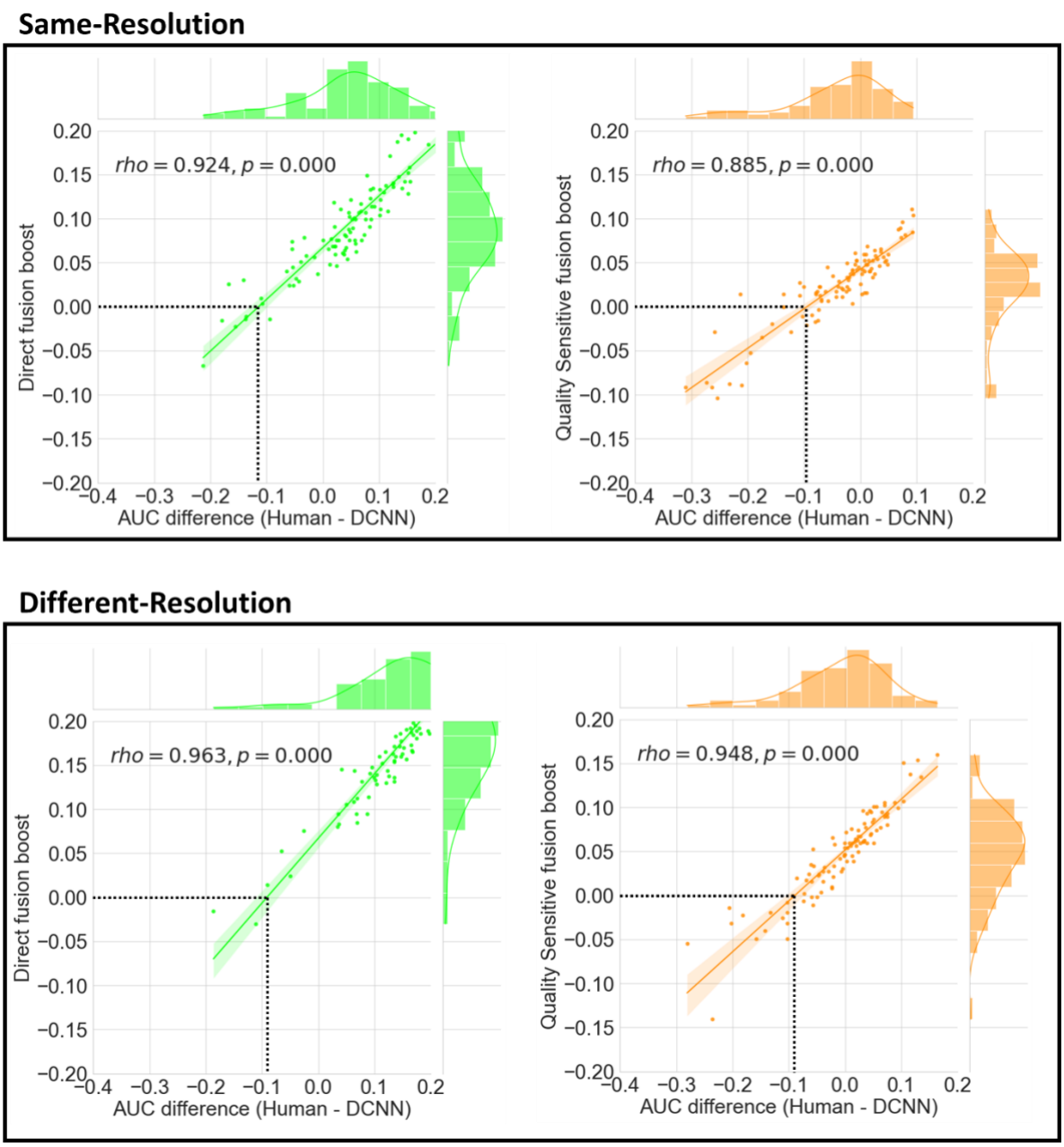


Figure B.2. Correlation between the difference between human and average DCNN performance (AUC) with the boost in performance in the two fusion metrics. At the top panel, we show the results for the Same-Resolution, and at the bottom, the Different-Resolution condition. We calculate the boost in performance by subtracting the resulting Fusion AUC from the DCNN AUC.

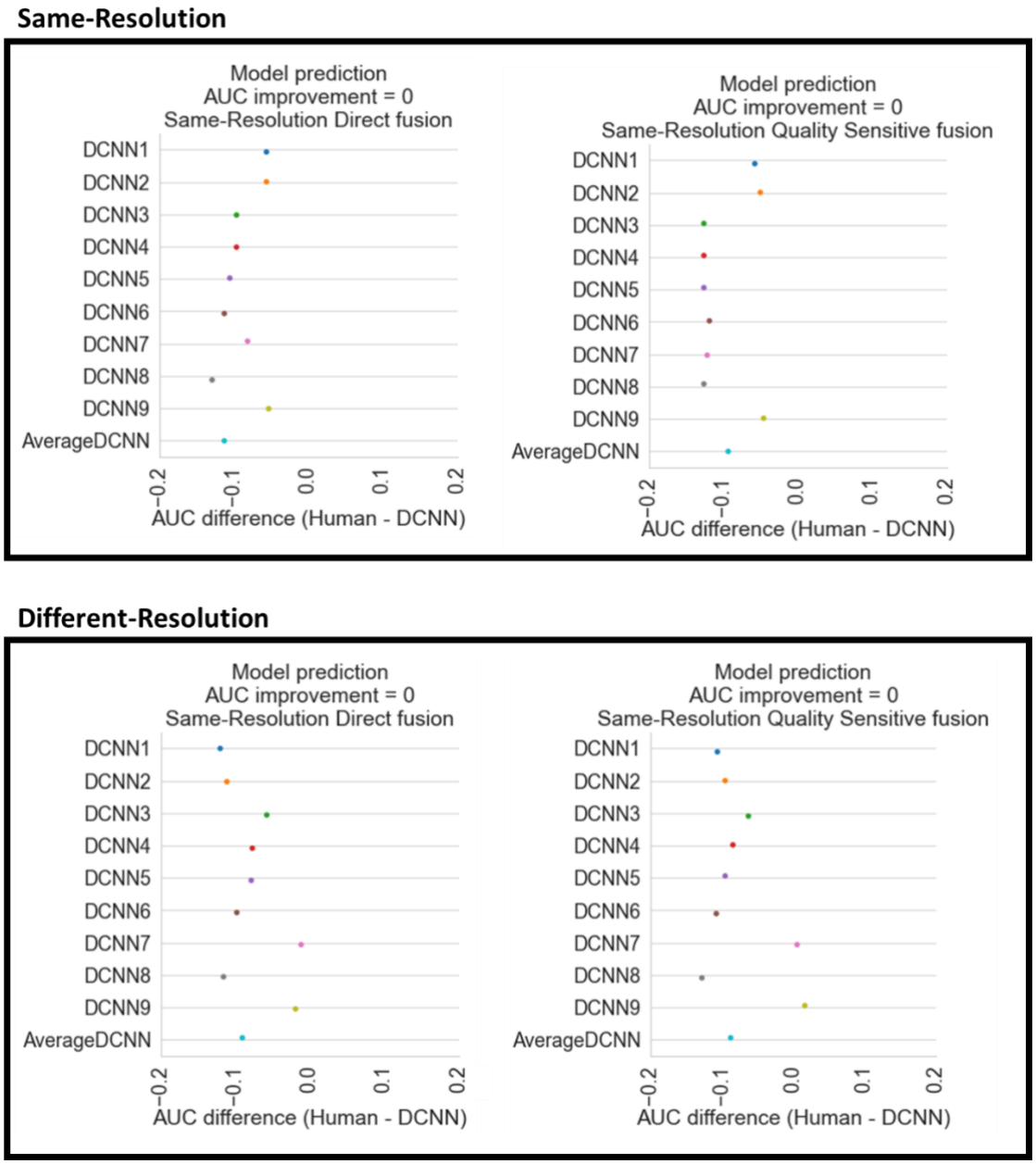


Figure B.3. Graphs showing the predicted accuracy that humans should show to benefit the fusion concerning the DCNNs used in this study.

Figure B.4 shows that the fusion of humans that disagree with DCNNs help to improve overall accuracy. We divided Figure B.4 into two panels. The top panel represents the findings for the Quality Sensitive fusion in the Same-Resolution, and the bottom panel, the Different-Resolution condition. We replicated the Quality Sensitive fusion correlation analysis from Figure B.2 on the left for each panel in Figure B.4. However, we used the linear trending regression to predict the accuracy position where humans start to benefit the fusion. This regression allowed us to determine the exact position where the model predicts the humans that would improve the fusion, dividing the data into expected ***positive*** (green

circles) and ***negative*** (red ‘x') boosts in performance. In addition, we calculated the residuals (i.e. the vertical distance) between each data coordinate against the predicted model. As aforementioned, we expect that higher agreement would lower the fusion effect. That is, a higher agreement should show lower residual values. The graphs on the right side of Figure B.4 show correlations between the agreement of humans and the average DCNN against the found residuals in the Quality Sensitive fusion. We found that the overall correlation was negative and significant for both the Same-Resolution [rho(98)= -0.43, p< 0.001] and Different Resolution [rho(98)= -0.60, p< 0.001] conditions. We later separated this analysis into humans who showed positive and negative boosts predicted by the linear model. This analysis showed that humans that positively boosted DCNNs showed larger fusion effects for the Same-Resolution [rho(98)= -0.71, p< 0.001] and Different-Resolution [rho(98)= -0.77, p< 0.001] conditions in comparison with humans that negatively boosted the DCNN in the Same-Resolution [rho(98)= -0.32, p< 0.001] and Different-Resolution [rho(98)= -0.61, p< 0.001]. Therefore, this result suggests that humans who show higher accuracy and rank their decisions differently than DCNNs will benefit the fusion even more.

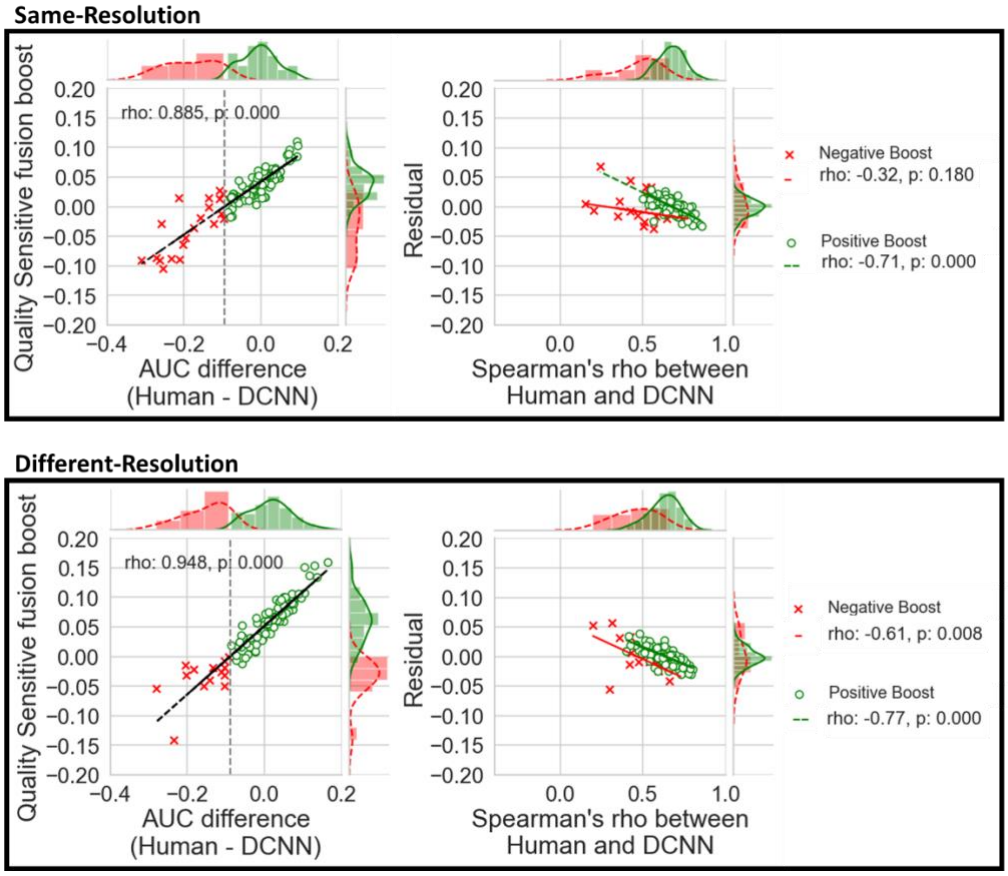


Figure B.4. The graphs show that humans who disagree with DCNNs' decisions improve fusion effects. At the top panel, we show the results for the Same-Resolution, and at the bottom, the Different-Resolution condition. On the left, the graphs show the correlation between human and average DCNN performance (AUC) with the boost in performance in the Quality Sensitive fusion metrics. On the right, we show the correlation between the residuals found in the graphs on the left with the agreement (Spearman's rho) between humans and DCNNs.

Figure B.5 illustrates the accuracy of the Direct fusion compared to humans and DCNNs, and it shows that the fusion effect is related to human accuracy. We divided Figure B.5 into two panels, where the top panel shows the results for the Same-Resolution and the bottom for the Different-Resolution conditions. Each panel shows the fusion effect of human participants and DCNNs in three different groups ranked by human performance: the top 50, all participants, and the bottom 50. Figure B.5 shows that fusing humans with DCNNs significantly improves - or decreases - DCNN's performance. Notably, the accuracy of the human participating in this fusion determines the outcome of this fusion. Figure B.6 shows similar results for the Quality Sensitive fusion. However, because we manipulated similarity scores differently in this fusion, DCNNs showed improved accuracy compared to the Direct fusion.

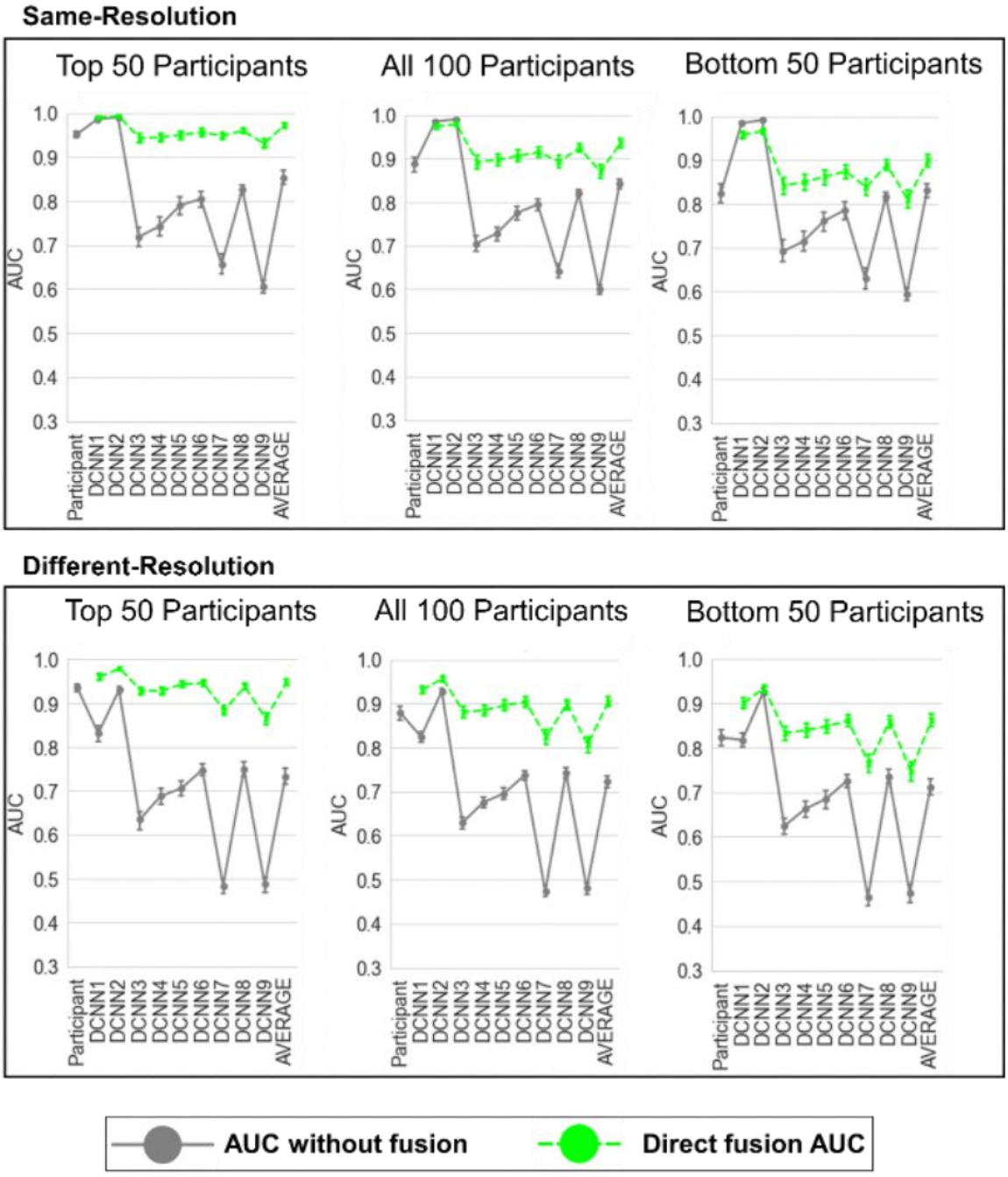


Figure B.5. Graphs showing the accuracy of humans, DCNNs, and the Direct fusion. At the top, we show the results for the top 50, all, and bottom 50 participants ranked by human accuracy in the Same-Resolution condition. At the bottom, we replicate the results for the Different-Resolution condition.

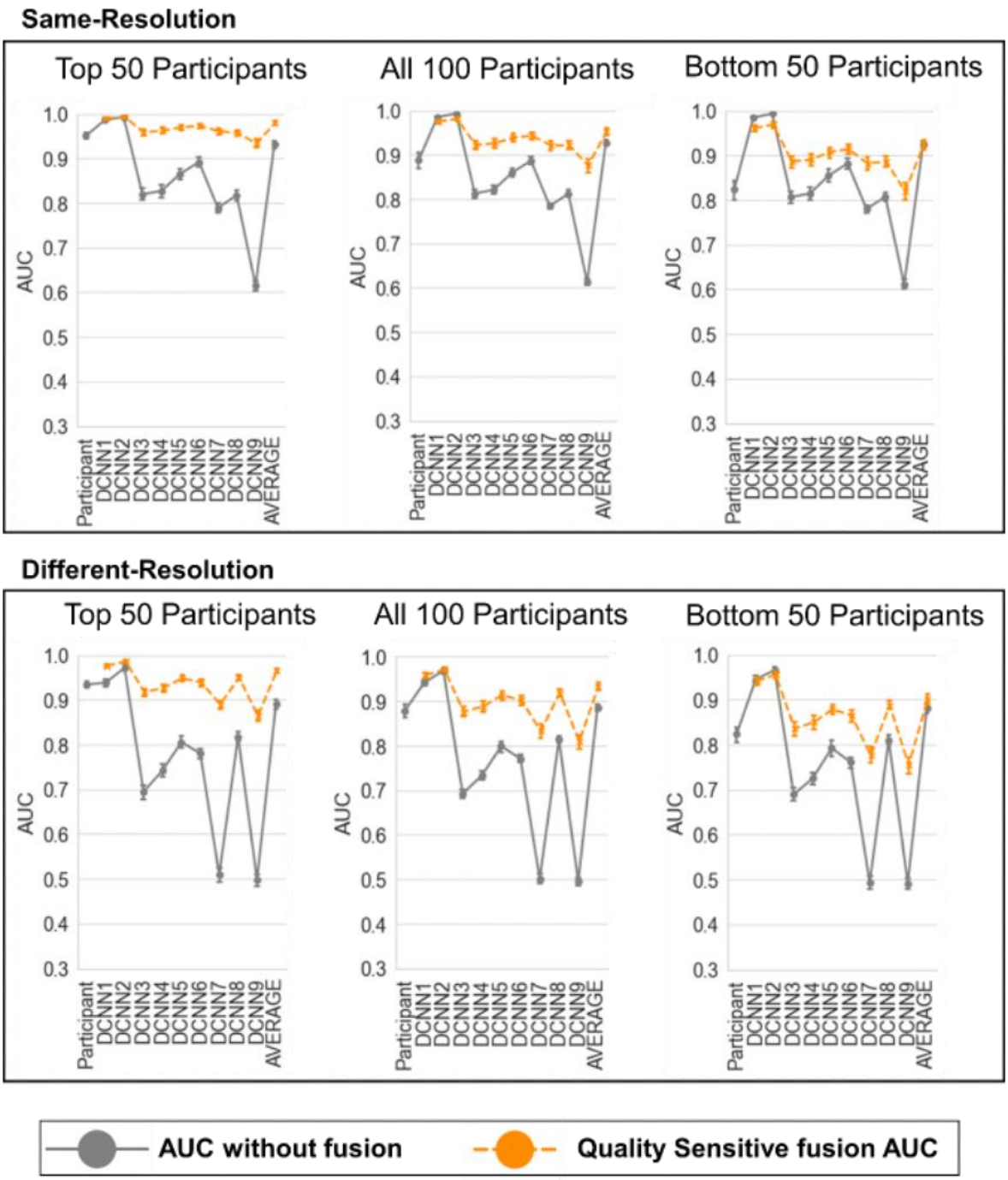


Figure B.6. Graphs showing the accuracy of humans, DCNNs, and the Quality Sensitive fusion. At the top, we show the results for the top 50, all, and bottom 50 participants ranked by human accuracy in the Same-Resolution condition. At the bottom, we replicate the results for the Different-Resolution condition.

APPENDIX C

Experiment 2 - Results for trial-level data

We performed a PCA on the trial-level heatmap data. In this analysis, the average heatmap receives a loading score of zero for each PC, and a zero-mean normal distribution represents participants' loading scores for each trial. Therefore, for every PC, some trials received a negative loading score (i.e. to the left side of the average), and some received a positive loading score (i.e. to the right side of the average). We show the visual reconstruction of the first five Principal Components (PCs) and how manipulating loading scores within PCs visually interacted with the average fixation heatmap in Figure C.1.

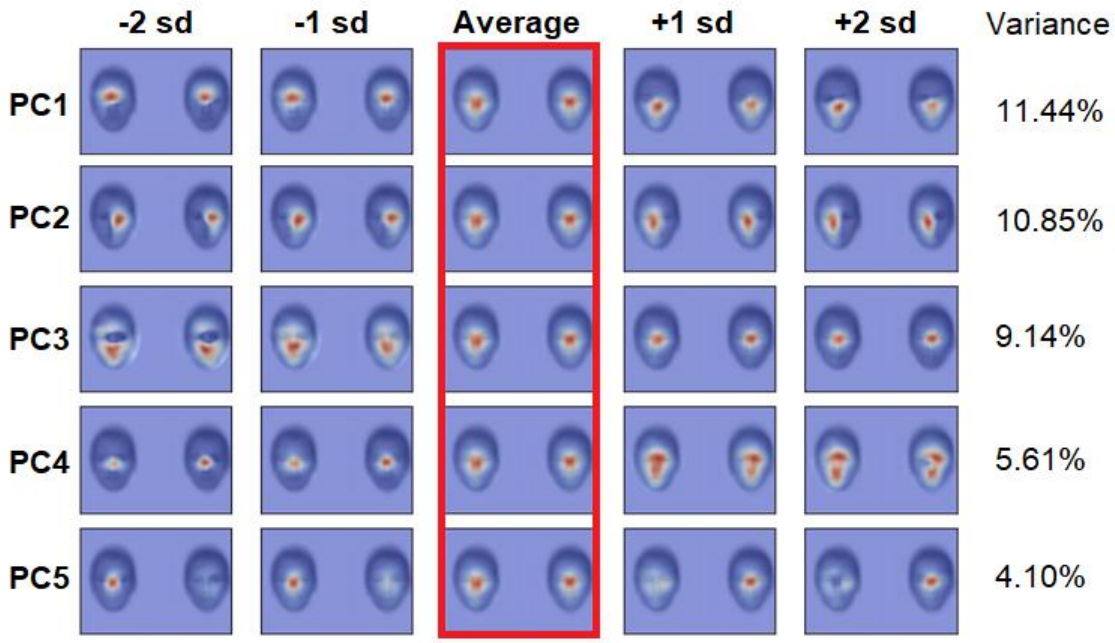


Figure C.1. Interaction of the average heatmap with the first five principal components. This figure shows PCs obtained using the trial-level heatmap data and their corresponding explained variances. In the PCA, the average heatmap receives a loading score of zero for each PC, and a zero-mean normal distribution represents participants' loading scores. Therefore, some trials received a negative loading score (i.e. to the left side of the average). And some received a positive loading score (i.e. to the right side of the average).

Figure C.1 visually described the five first PCs for the trial-level heatmap data, representing semantically explainable shifts in fixation patterns similar to those found in the main manuscript. So, to support the PC investigation in Figure C.1, we performed a

correlational study of PC loading scores with ROI fixations. See Table C.1. For correlations of PC loading scores and ROI fixations against face recognition ability, see Table C.2.

Correlation Matrix - Principal Components and ROIs - Trial-level data

		Left eyes		Right eyes		Between eyes		Noses		Mouths		Other features	
PC1	Spearman's rho	-0.34	***	0.009		-0.29	***	0.691	***	0.486	***	-0.17	***
	p-value	< .001		0.702		< .001		< .001		< .001		< .001	
PC2	Spearman's rho	0.536	***	-0.55	***	0.025		0.288	***	0.15	***	-0.1	***
	p-value	< .001		< .001		0.301		< .001		< .001		< .001	
PC3	Spearman's rho	0.347	***	0.537	***	0.717	***	0.102	***	-0.18	***	-0.42	***
	p-value	< .001		< .001		< .001		< .001		< .001		< .001	
PC4	Spearman's rho	-0.08	***	0.144	***	-0.17	***	-0.21	***	0.52	***	0.246	***
	p-value	< .001		< .001		< .001		< .001		< .001		< .001	
PC5	Spearman's rho	0.063	**	0.089	***	-0.04		-0.15	***	-0.02		0.092	***
	p-value	0.008		< .001		0.139		< .001		0.507		< .001	

Note. * p < .05, ** p < .01, *** p < .001

Table 4.1. Correlational analysis of Principal Component loading scores with fixations in ROI.

Correlation Matrix
PCs vs GFMT Score

		GFMT Score	
PC1	Spearman's rho	0.049	*
	p-value	0.041	
PC2	Spearman's rho	-0.159	***
	p-value	< .001	
PC3	Spearman's rho	0.219	***
	p-value	< .001	
PC4	Spearman's rho	0.268	***
	p-value	< .001	
PC5	Spearman's rho	0.056	*
	p-value	0.019	

Note. * p < .05, ** p < .01, *** p < .001

Correlation Matrix
ROIs vs GFMT Score

		GFMT Score	
Left eyes	Spearman's rho	-0.014	
	p-value	0.567	
Right eyes	Spearman's rho	0.317	***
	p-value	< .001	
Between eyes	Spearman's rho	0.089	***
	p-value	< .001	
Noses	Spearman's rho	-0.076	**
	p-value	0.002	
Mouths	Spearman's rho	0.107	***
	p-value	< .001	
Other features	Spearman's rho	0.034	
	p-value	0.159	

Table C.2. Tables showing how Principal Components (PCs) and fixations on ROIs correlated with face-matching ability measured by the stimuli (GFMT: Burton et al., 2010). We show Spearman's correlation between the five first PCs against GFMT score on the left table. And show Spearman's correlation between fixations in ROIs against GFMT score on the right table.

The MLDA assigned participants with MLDA Scores within its resulting eigenvector. The t-test [t(1757)= 36.22, p< 0.001, Cohen's d= 2.06] shows that MLDA scores significantly separated the two groups of participants. We show the MLDA eigenvector at the top panel of Figure C.2 as a heatmap image. We could not subjectively interpret this heatmap image.

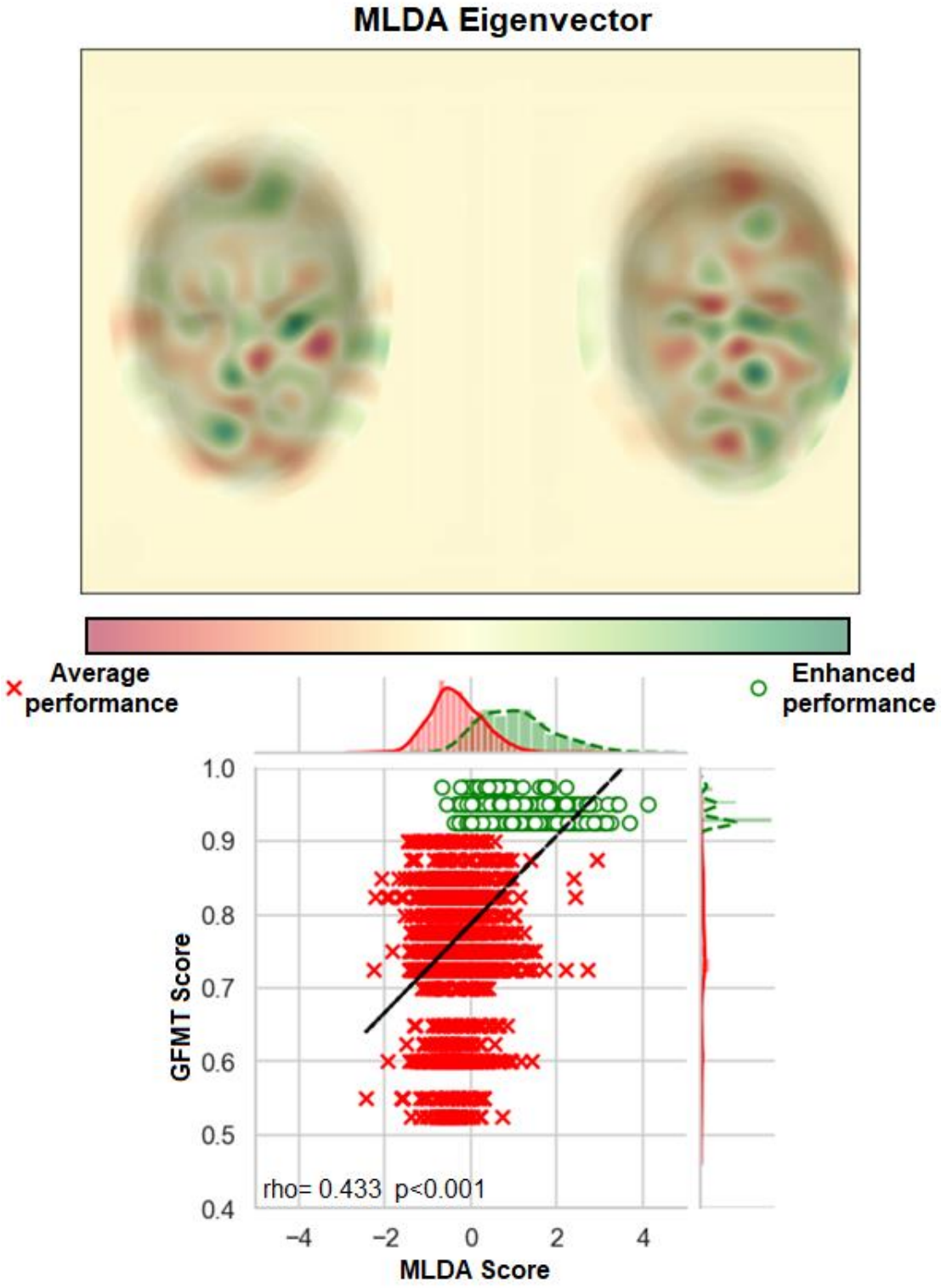


Figure C.2. MLDA eigenvector and the correlation of MLDA scores with face-matching performance. At the top, we illustrate the MLDA eigenvector - as a heatmap image. The MLDA eigenvector describes what best differentiates the fixation patterns between two categories of participants. At the bottom, we show the MLDA Score distribution amongst participants, the average score for each group (vertical dashed lines), and the correlation between MLDA scores against face-matching ability.

To further investigate the MLDA scores distribution, we performed a correlational analysis of participants' MLDA scores against the five first PC loading scores and ROI fixations. See Table C.3.

Correlation Matrix

		MLDA Score	
PC1	Spearman's rho	0.233	***
	p-value	< .001	
PC2	Spearman's rho	-0.33	***
	p-value	< .001	
PC3	Spearman's rho	0.479	***
	p-value	< .001	
PC4	Spearman's rho	0.117	***
	p-value	< .001	
PC5	Spearman's rho	-0	
	p-value	0.881	

Correlation Matrix

		MLDA Score	
Left eyes	Spearman's rho	-0.05	*
	p-value	0.037	
Right eyes	Spearman's rho	0.593	***
	p-value	< .001	
Between eyes	Spearman's rho	0.294	***
	p-value	< .001	
Noses	Spearman's rho	-0.014	
	p-value	0.55	
Mouths	Spearman's rho	0.141	***
	p-value	< .001	
Other features	Spearman's rho	-0.271	***
	p-value	< .001	

Note. * p < .05, ** p < .01, *** p < .001

Table C.3. Tables show how Principal Components (PCs) and fixations on ROIs correlated with MLDA Score. We show Spearman's correlation between the five first PCs against MLDA Scores on the left table and fixations in ROIs against MLDA scores on the right table.

The trial-level data allows us to investigate the fixation pattern's stability of participants processing the stimuli. That is, we could use the PCA – or MLDA - to explain the different fixation patterns within participants. Figure C.3 shows the dispersion of PC1 loading and MLDA scores across participants. Interestingly, visual inspection of Figure C.3 shows substantial inter-trial variability. This variability could explain that participants do not possess a single fixation pattern strategy to process face-matching trials. We found a similar pattern of results for the remaining PCs.

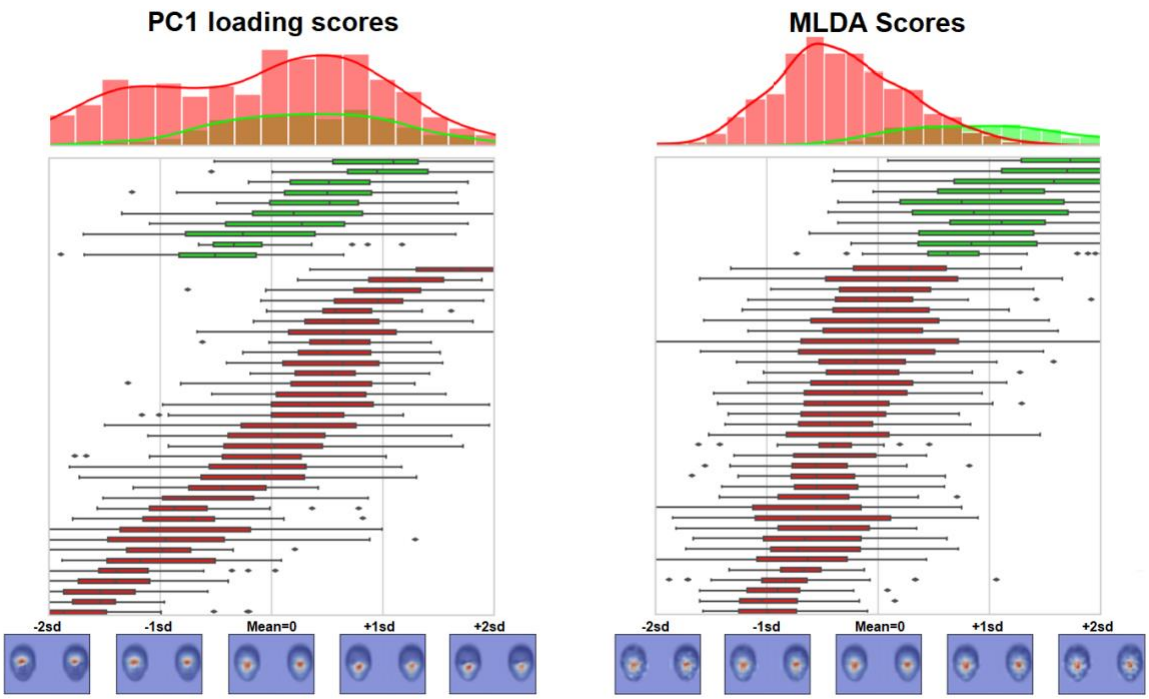


Figure C.3. Intra-individual variation in loading scores across GFMT trials for Principal Component 1 (left) and MLDA scores (right). Although only PC1 is shown here, variation is similar in magnitude for all PCs. The red bars stand for participants in the Average performance group. The green bars stand for participants in the Enhanced performance group. Participants' data are ordered from top to bottom based on the mean loading scores of each group, but show extremely large intra-individual variation in all participants.

We investigated whether the amount of exploration participants engaged correlated with performance. For that, we calculated the Gini coefficient (Lorenz, 1905) for each trial-level heatmap and correlated these coefficients with face processing ability, PCs, MLDA, and time elapsed per trial. As a reminder, lower Gini coefficients represent higher exploration. We found that this measure of exploration significantly correlated with PCs, MLDA Scores, and face-matching ability. See Table C.4.

Correlation Matrix Gini coefficient

		Gini coefficient	
GFMT Score	Spearman's rho	-0.335	***
	p-value	< .001	
PC1	Spearman's rho	-0.199	***
	p-value	< .001	
PC2	Spearman's rho	-0.193	***
	p-value	< .001	
PC3	Spearman's rho	-0.201	***
	p-value	< .001	
PC4	Spearman's rho	-0.623	***
	p-value	< .001	
PC5	Spearman's rho	-0.131	***
	p-value	< .001	
MLDA Score	Spearman's rho	-0.228	***
	p-value	< .001	
Time Elapsed	Spearman's rho	-0.502	***
	p-value	<.001	

Note. * p < .05, ** p < .01, *** p < .001

Table C.4. *Correlation between trial-level Gini coefficients with GFMT Scores, Principal Component loading scores, MLDA Scores, and Time spent on each trial.*

Experiment 3A

Aperture size determination

We based on the work of Papinutto and colleagues (2017) to determine the aperture sizes we used in Experiment 3. In their work, they show a data-driven reconstruction of the Facespan based on the convolution of a retinal filter (Targino Da Costa & Do, 2014),

calculation of the SIMilarity index (Wang, Bovik, Sheikh & Simoncelli, 2004), and the pixel-test (Cahuvin, Worsley, Schyns, Arguin & Gosselin, 2005; Random field Theory). Their analysis shows that a 17º Gaussian aperture corresponded to 7º (45%) of the total face information being available. Therefore, assuming linearity, the five apertures (5º, 10º, 15º, 20º, and 25º) correspond to 2°,4°,6°,8°, and 10° of information available for every fixation. Here, we report these apertures regarding the percentage of face information available for each fixation, corresponding to 12%, 24%, 36%, 48%, and 60% of facial information available for every fixation. Natural view (NV) provided 100% of the face. See video demonstration at: https://osf.io/xtjzh/.

PCA – Learning and Recognition phases

We show the PCA navigation for the five first PCs in Figure C.4. In Figure C.4, we divided the PCA analysis into Learning and Recognition phases. Inspection of this Figure reveals that the first 5 PCs for each phase were similar as they visually represent similar changes in fixation patterns.

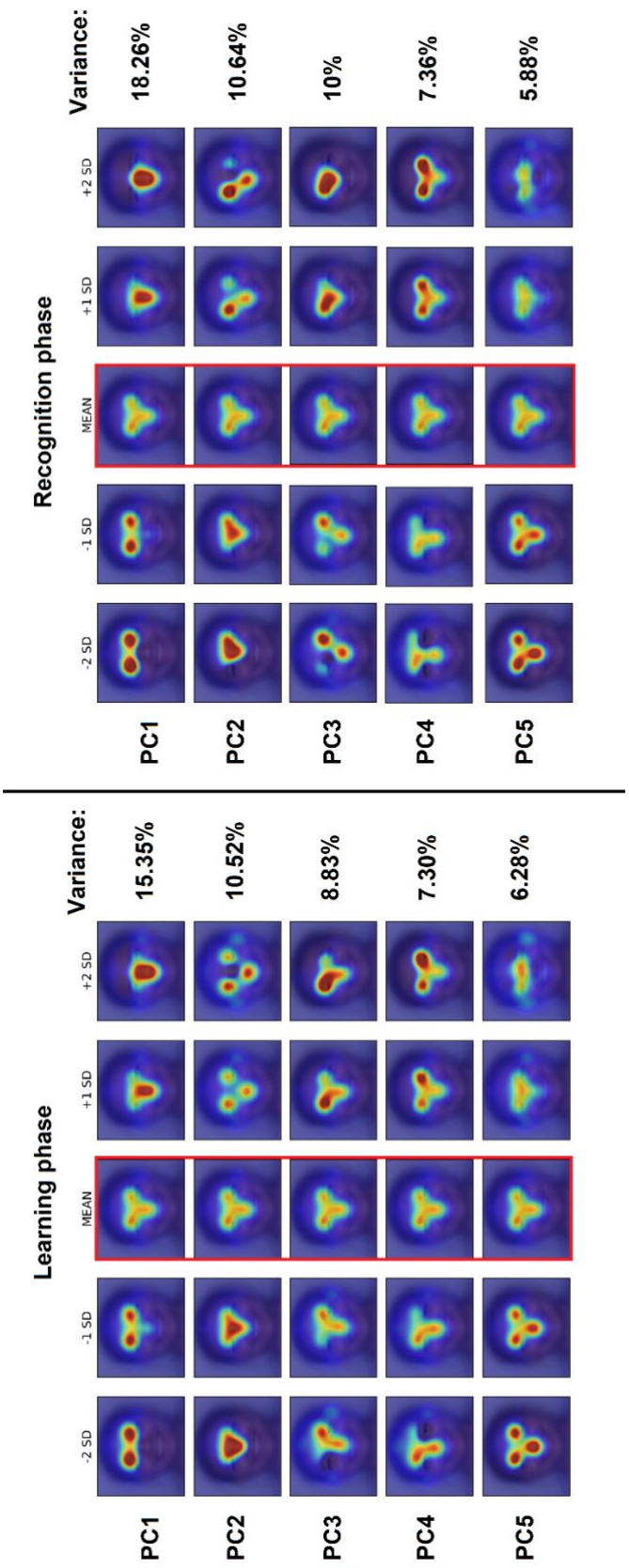


Figure C.4. PCA navigation for Learning and Recognition phases.

We analysed PC loadings using a Linear Mixed Model. For this analysis, we used the function: Loading ~ 1 + group + phase + aperture size + group:phase + group:aperture size + phase:aperture seize + group:phase:aperture size + (1 | image) + (1 | participant). See Tables C.5 – C.9.

Investigating PC1, we observed a significant interaction between phases and group [b = -.101, CI = [-.154, -.048], t(11511.6) = 3.72, p < .001]. This significant result shows that PC1 loading scores differentiate typical viewers' fixation patterns from super recognisers'. In this analysis, the simple effects illustrate a larger difference between groups during face learning [b=0.225] than recognition [b=0.125]. Interestingly, the significant two-way interaction between group and aperture size [b=-0.05, CI =[-0.08,-0.03], t(10436.3)=-3.79, p<0.001], and between group and phase [b=-0.11, CI =[-0.15,-0.05], t(11511.6)=-3.72, p<0.001] reveal that super recognisers differed from typical viewers in the sampled information described by PC1 across aperture size and phase. However, the not significant three-way interaction of group, phase and aperture condition [b=-0.01, CI =[-0.07, 0.04], t(11516.6)= -0.46, p= 0.642] reveal that these differences in the information sampled between groups remained consistent in all aperture conditions and phases. See Figure 4.10B. Visual inspection of Figure 4.10B reveals a tendency of typical viewers to show a more negative loading score compared to super recognisers. According to our subjective interpretation of PC1, a negative loading score relates to preferences to observe the eye region more than other facial parts. Therefore, our analysis shows that typical viewers tended to observe the eye region to memorise faces more than super-recognisers across aperture conditions. In contrast, super-recognisers preferred to investigate the central features of the face, not showing too much importance to the eye region. However, we also found a significant main effect of aperture size [b = 0.07, CI = [0.05, 0.08], t(2808.8) = 7.97, p < .001], showing that for smaller aperture sizes, both typical viewers and super-recognisers approached the eye region (i.e. a more negative PC1 loading score).

Linear mixed model results for PC1 loading

			95% Confidence Interval				
Fixed Effects	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	-0.00879	0.09342	-0.1919	0.1743	59.6	-0.0941	0.925
Group	0.17504	0.18569	-0.1889	0.539	58.1	0.9427	0.35
Phase	0.00521	0.01426	-0.0227	0.0332	11111.1	0.3655	0.715
Aperture	0.07057	0.00886	0.0532	0.0879	2808.8	7.9646	< .001
Group ✻ Phase	-0.10087	0.02713	-0.154	-0.0477	11511.6	-3.7182	< .001
Group ✻ Aperture	-0.05356	0.01411	-0.0812	-0.0259	10436.3	-3.7954	< .001
Phase ✻ Aperture	0.0773	0.01405	0.0498	0.1048	11214.9	5.5034	< .001
Group ✻ Phase ✻ Aperture	-0.0125	0.02688	-0.0652	0.0402	11516.6	-0.4649	0.642

Table C.5. Linear mixed model results for PC1 loading.

Linear mixed model results for PC2 loading

			95% Confidence Interval				
Names	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	-6.84e−4	0.0717	-0.1411	0.13977	61.2	-0.00955	0.992
Group	0.0517	0.1415	-0.2257	0.32913	58.2	0.36541	0.716
Phase	-0.0158	0.0165	-0.0481	0.01657	11024.3	-0.956	0.339
Aperture	-0.2077	0.0101	-0.2276	-0.18782	2779.4	-20.46696	< .001
Group ✻ Phase	-0.0218	0.0314	-0.0835	0.03982	11534.1	-0.69391	0.488
Group ✻ Aperture	-0.0426	0.0163	-0.0746	-0.01059	10399	-2.60909	0.009
Phase ✻ Aperture	-0.0342	0.0163	-0.066	-0.0023	11143.8	-2.10152	0.036
Group ✻ Phase ✻ Aperture	-0.0478	0.0312	-0.1089	0.01324	11537.3	-1.535	0.125

Table C.6. Linear mixed model results for PC2 loading.

Linear mixed model results for PC3 loading

Names	Estimate	SE	95% Confidence Interval		df	t	p
			Lower	Upper			
(Intercept)	-0.00513	0.06775	-0.1379	0.1277	59.8	-0.0757	0.94
Group	0.05951	0.13456	-0.2042	0.3232	58.2	0.4422	0.66
Phase	0.01056	0.01679	-0.0223	0.0435	9159.6	0.629	0.529
Aperture	0.18005	0.00966	0.1611	0.199	1143.9	18.6439	< .001
Group ✻ Phase	-0.04411	0.03243	-0.1077	0.0195	11577.6	-1.36	0.174
Group ✻ Aperture	0.00163	0.01655	-0.0308	0.0341	8818.7	0.0984	0.922
Phase ✻ Aperture	0.207	0.01656	0.1745	0.2395	9610.1	12.4969	< .001
Group ✻ Phase ✻ Aperture	0.04043	0.03213	-0.0225	0.1034	11576.8	1.2582	0.208

Table C.7. Linear mixed model results for PC3 loading.

Linear mixed model results for PC4 loading

Names	Estimate	SE	95% Confidence Interval		df	t	p
			Lower	Upper			
(Intercept)	-0.01121	0.0676	-0.14369	0.1213	62.3	-0.16583	0.869
Group	0.27065	0.133	0.01004	0.5313	58.3	2.03549	0.046
Phase	-0.00147	0.0172	-0.03517	0.0322	11108	-0.08548	0.932
Aperture	0.01943	0.0106	-0.00144	0.0403	2925.7	1.82513	0.068
Group ✻ Phase	0.14967	0.0327	0.08551	0.2138	11524.1	4.57261	< .001
Group ✻ Aperture	-0.05797	0.017	-0.09131	-0.0246	10481.9	-3.40777	< .001
Phase ✻ Aperture	-7.13e−5	0.0169	-0.03327	0.0331	11211.2	-0.00421	0.997
Group ✻ Phase ✻ Aperture	0.01142	0.0324	-0.05214	0.075	11528.3	0.35215	0.725

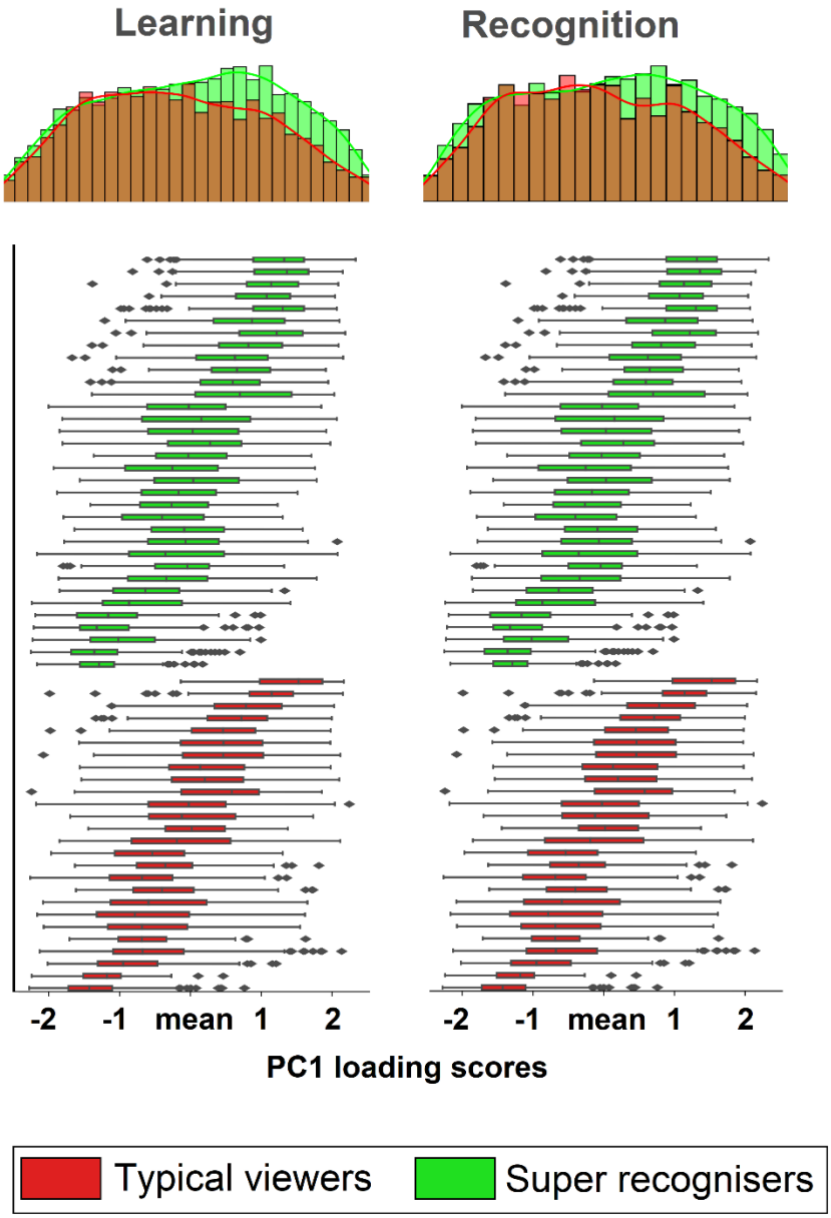
Table C.8. Linear mixed model results for PC4 loading.

Linear mixed model results for PC5 loading

Names	Estimate	SE	95% Confidence Interval		df	t	p
			Lower	Upper			
(Intercept)	0.00292	0.0678	-0.12989	0.1357	62.8	0.043	0.966
Group	-0.16011	0.133	-0.42076	0.1005	58.3	-1.204	0.233
Phase	0.02561	0.0174	-0.00841	0.0596	11253.7	1.4752	0.14
Aperture	-0.05374	0.0108	-0.07499	-0.0325	3399.5	-4.9543	< .001
Group ✻ Phase	-0.12362	0.033	-0.18827	-0.059	11514	-3.7477	< .001
Group ✻ Aperture	0.07301	0.0172	0.03933	0.1067	10693.6	4.2487	< .001
Phase ✻ Aperture	0.04549	0.0171	0.01198	0.079	11327.6	2.661	0.008
Group ✻ Phase ✻ Aperture	-0.05022	0.0327	-0.11427	0.0138	11519.1	-1.5366	0.124

Table C.9. Linear mixed model results for PC5 loading.

We investigated the stability of fixation patterns measured by PC1 scores. See Figure C.5. To illustrate the order of magnitude of this variation, we found that the average range of participants' PC1 scores (*range = max - min*) in the learning phase was of 3.24 standard deviations (Typical viewers= 3.33; Super-recognisers= 3.17; t(58)=1.13, p=.262), and in the recognition phase was of 3.09 standard deviations (Typical viewers= 3.13; Super-recognisers= 3.06; t(58)=0.58, p=.563). This vast range of PC1 scores per participants suggests that individual trials influence participants to engage in particular fixation patterns. In addition, we found a great relationship between PC1 loading scores for learning and recognition phases. This result suggests that despite the large intra-individual differences, participants showed significant consistency between their fixation patterns across phases ($r(58) = 0.95$, $p < 0.001$). Curiously, we also found stability in the standard deviation of PC1 loading scores between phases ($r(58) = 0.69$, $p < 0.001$), raising evidence that the fixation pattern differences within participants also show a stable trait.



C.5. Trial-level variability in PC1 loading scores across participants. The boxes show the interquartile range, and the whiskers show the minimum and maximum of the distribution.

Gaze dispersal Analysis (all apertures)

We used linear mixed model analysis to investigate the exploration of participants processing the stimuli using Gini coefficients. For the model using the Gini Coefficient, we set participants' intercept as a random effect and group, aperture size, and phase as fixed effects. Linear mixed models reveal a significant main effect of phase [b= .018, CI= [.016, .019], t(11304.4)= 22.52, p< 0.001] and aperture size [b= .002, CI= [.002, .004], t(3865)=

5.35, p< 0.001]. In addition, we found that the three-way interaction between group, phase, and aperture size [b= .004, CI= [.001, .007], t(11503.5)= 2.82, p= 0.005] was also significant. Investigating further, simple effects found lower Gini coefficients (i.e. more exploration) for super-recognisers at larger aperture sizes compared to typical viewers [b = -.013, t(63.6) = -1.73, p= 0.089] and this difference decreased with smaller aperture sizes [b = -.003, t(63.7) = -0.41, p= 0.683]. We replicated this analysis only considering the recognition phase and found that the difference between groups was smaller in magnitude (b's between -0.001 to -0.002). This analysis shows that super recognisers explore faces more than typical viewers. However, similar to the analysis in the main document, this pattern was more prominent during the learning phase and in larger aperture conditions.

PCA – Learning and Recognition (Only NV)

We show the PCA navigation for the five first PCs in Figure C.6. In Figure C.6, we divided the PCA analysis into Learning and Recognition phases. Inspection of this Figure reveals that the first 5 PCs for each phase were similar as they visually represent similar changes in fixation patterns.

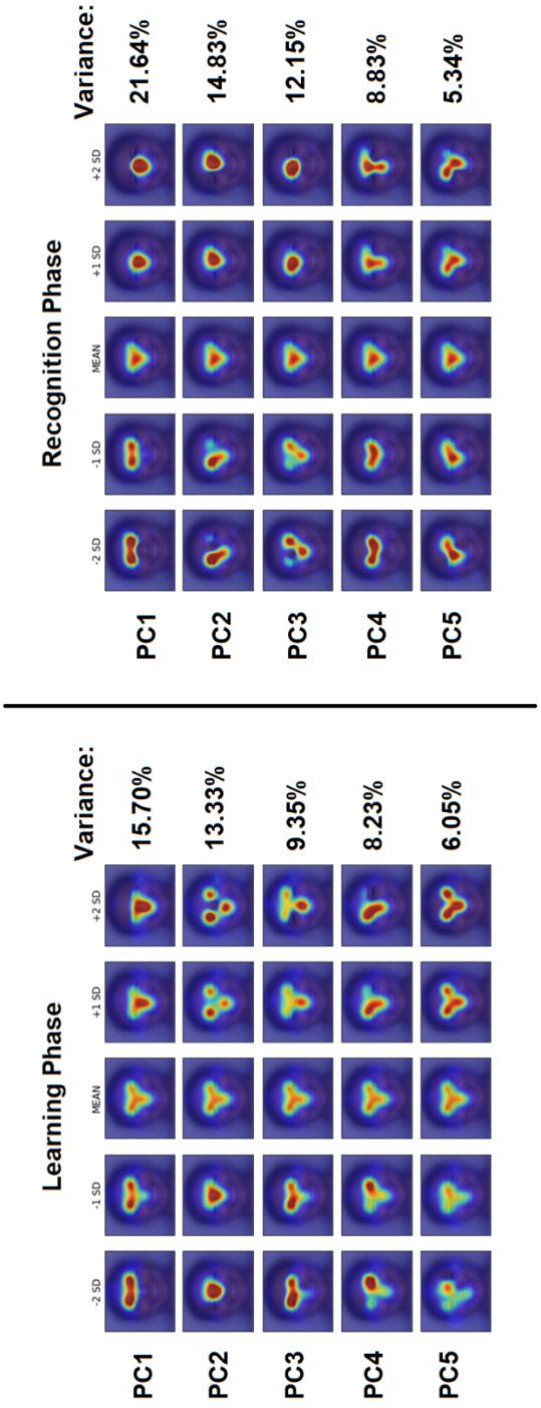


Figure C.6. PCA navigation for Learning and Recognition phases.

We analysed PC loadings using a Linear Mixed Model. For this analysis, we used the function: Loading ~ 1 + Phase + Group + Phase:Group + (1| ID). See Tables C.10 – C.14.

Linear Mixed Model results for PC1

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.0195	0.0959	-0.2074	0.1684	58.1	-0.203	0.84
Phase1	Recognition - Learning	-0.0289	0.0338	-0.0951	0.0374	1745.9	-0.854	0.393
Group1	SuperRecogniser - Typical Viewer	0.0943	0.1917	-0.2815	0.4701	58.1	0.492	0.625
Phase1 ✻ Group1	Recognition - Learning ✻ SuperRecogniser - Typical Viewer	-0.302	0.0676	-0.4345	-0.1695	1745.9	-4.466	< .001

Table C.10. Linear mixed model results for PC1 loading.

Linear Mixed Model results for PC2

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-7.02e−4	0.0393	-0.0777	0.0762	58.6	-0.0179	0.986
Phase1	Recognition - Learning	-0.00238	0.0471	-0.0948	0.09	1760.6	-0.0506	0.96
Group1	SuperRecogniser - Typical Viewer	0.07549	0.0785	-0.0784	0.2294	58.6	0.9614	0.34
Phase1 ✻ Group1	Recognition - Learning ✻ SuperRecogniser - Typical Viewer	0.1219	0.0943	-0.0629	0.3067	1760.6	1.2927	0.196

Table C.11. Linear mixed model results for PC2 loading.

Linear Mixed Model results for PC3

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.00525	0.0476	-0.0985	0.088	59.1	-0.1102	0.913
Phase1	Recognition - Learning	-0.0034	0.0461	-0.0938	0.087	1756.3	-0.0738	0.941
Group1	SuperRecogniser - Typical Viewer	-0.00963	0.0952	-0.1962	0.1769	59.1	-0.1012	0.92
Phase1 ✻ Group1	Recognition - Learning ✻ SuperRecogniser - Typical Viewer	0.00847	0.0923	-0.1724	0.1893	1756.3	0.0918	0.927

Table C.12. Linear mixed model results for PC3 loading.

Linear Mixed Model results for PC4

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	0.0098	0.0638	-0.1152	0.1348	58.8	0.154	0.878
Phase1	Recognition - Learning	0.03336	0.0428	-0.0505	0.1172	1750.7	0.78	0.436
Group1	SuperRecogniser - Typical Viewer	-0.18201	0.1275	-0.432	0.0679	58.8	-1.427	0.159
Phase1 ✻ Group1	Recognition - Learning ✻ SuperRecogniser - Typical Viewer	-0.25044	0.0856	-0.4181	-0.0828	1750.7	-2.927	0.003

Table C.13. Linear mixed model results for PC4 loading.

Linear Mixed Model results for PC5

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.0161	0.0636	-0.1407	0.1085	58	-0.253	0.801
Phase1	Recognition - Learning	-0.0198	0.0434	-0.105	0.0653	1750.1	-0.456	0.648
Group1	SuperRecogniser - Typical Viewer	0.0891	0.1271	-0.1601	0.3383	58	0.701	0.486
Phase1 ✻ Group1	Recognition - Learning ✻ SuperRecogniser - Typical Viewer	0.2065	0.0869	0.0362	0.3768	1750.1	2.377	0.018

Table C.14. Linear mixed model results for PC5 loading.

Experiment 3B

PCA – Learning and Recognition

We analysed PC loadings using a Linear Mixed Model. For this analysis, we used the function: Loading ~ 1 + CFMTp + Phase + Aperture + Phase:CFMTp + CFMTp:Aperture + Phase:CFMTp:Aperture + (1| ID) + (1 | Image). See Tables C.15 – C.19.

Linear Mixed Model results for PC1

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.00751	0.10606	-0.21539	0.2004	45.2	-0.0708	0.944
CFMTp	CFMTp	0.02062	0.10433	-0.18386	0.2251	43	0.1976	0.844
Phase1	Recognition - Learning	0.0102	0.01687	-0.02287	0.0433	8521.1	0.6047	0.545
Aperture	Aperture	0.10744	0.01386	0.08028	0.1346	653.2	7.7538	< .001
CFMTp ✻ Phase1	CFMTp ✻ Recognition - Learning	-0.05685	0.01505	-0.08634	-0.0274	9155	-3.7775	< .001
CFMTp ✻ Aperture	CFMTp ✻ Aperture	0.0424	0.00751	0.02768	0.0571	9220.3	5.6454	< .001
Phase1 ✻ Aperture	Recognition - Learning ✻ Aperture	0.08964	0.01671	0.05689	0.1224	8575.1	5.3644	< .001
CFMTp ✻ Phase1 ✻ Aperture	CFMTp ✻ Recognition - Learning ✻ Aperture	0.03253	0.01499	0.00315	0.0619	9157.4	2.1704	0.03

Table C.15. Linear mixed model results for PC1 loading.

Linear Mixed Model results for PC2

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	-0.02001	0.06595	-0.1493	0.10924	61.5	-0.303	0.763
CFMTp	CFMTp	-0.00173	0.00466	-0.0109	0.00741	43.2	-0.37	0.713
Phase1	Recognition - Learning	-0.01484	0.02103	-0.0561	0.02639	9186.3	-0.705	0.481
Aperture	Aperture	-0.00639	0.00223	-0.0108	-0.00201	1099.8	-2.859	0.004
CFMTp ✻ Phase1	CFMTp ✻ Recognition - Learning	0.00183	0.00144	-9.97e−4	0.00466	9133.5	1.269	0.205
CFMTp ✻ Aperture	CFMTp ✻ Aperture	1.34E-04	8.41E-05	-3.11e−5	2.99E-04	9216.9	1.59	0.112
Phase1 ✻ Aperture	Recognition - Learning ✻ Aperture	-0.00825	0.00243	-0.013	-0.00349	9174.6	-3.395	< .001
CFMTp ✻ Phase1 ✻ Aperture	CFMTp ✻ Recognition - Learning ✻ Aperture	4.11E-05	1.68E-04	-2.88e−4	3.70E-04	9137.1	0.245	0.807

Table C.16. Linear mixed model results for PC2 loading.

Linear Mixed Model results for PC3

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	-0.00533	0.07234	-0.14712	0.13646	49.2	-0.0737	0.942
CFMTp	CFMTp	-0.00829	0.00542	-0.01891	0.00233	43.1	-1.5295	0.133
Phase1	Recognition - Learning	0.00478	0.02011	-0.03463	0.04419	8350.1	0.2375	0.812
Aperture	Aperture	-0.01785	0.00188	-0.02155	-0.01416	633.1	-9.4781	< .001
CFMTp ✻ Phase1	CFMTp ✻ Recognition - Learning	-0.00306	0.0014	-0.00581	-3.19e−4	9168	-2.1879	0.029
CFMTp ✻ Aperture	CFMTp ✻ Aperture	-4.36e−5	8.15E-05	-2.03e−4	1.16E-04	9226.1	-0.5349	0.593
Phase1 ✻ Aperture	Recognition - Learning ✻ Aperture	-0.01706	0.00232	-0.02162	-0.01251	8428.6	-7.3444	< .001
CFMTp ✻ Phase1 ✻ Aperture	CFMTp ✻ Recognition - Learning ✻ Aperture	-8.47e−5	1.63E-04	-4.04e−4	2.34E-04	9170	-0.521	0.602

Table C.17. Linear mixed model results for PC3 loading.

Linear Mixed Model results for PC4

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	0.00632	0.06981	-0.1305	0.14314	51.5	0.0905	0.928
CFMTp	CFMTp	0.01143	0.00517	0.0013	0.02157	43.1	2.2117	0.032
Phase1	Recognition - Learning	0.02116	0.02074	-0.01949	0.06181	8617.2	1.0202	0.308
Aperture	Aperture	-0.00354	0.002	-0.00747	3.85E-04	696.7	-1.7679	0.078
CFMTp ✻ Phase1	CFMTp ✻ Recognition - Learning	-0.00582	0.00144	-0.00863	-0.003	9159.1	-4.0447	< .001
CFMTp ✻ Aperture	CFMTp ✻ Aperture	-7.08e−4	8.37E-05	-8.72e−4	-5.44e−4	9223.6	-8.4598	< .001
Phase1 ✻ Aperture	Recognition - Learning ✻ Aperture	3.63E-04	0.0024	-0.00433	0.00506	8660.1	0.1514	0.88
CFMTp ✻ Phase1 ✻ Aperture	CFMTp ✻ Recognition - Learning ✻ Aperture	-3.43e−4	1.67E-04	-6.71e−4	-1.59e−5	9161.5	-2.0551	0.04

Table C.18. Linear mixed model results for PC4 loading.

Linear Mixed Model results for PC5

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.00767	0.05688	-0.11916	0.10381	54.2	-0.135	0.893
CFMTp	CFMTp	0.00469	0.00416	-0.00346	0.01284	43.3	1.127	0.266
Phase1	Recognition - Learning	0.03215	0.0219	-0.01077	0.07508	7965.1	1.468	0.142
Aperture	Aperture	-0.01056	0.002	-0.01447	-0.00664	530.4	-5.284	< .001
CFMTp ✻ Phase1	CFMTp ✻ Recognition - Learning	2.06E-04	0.00153	-0.00279	0.00321	9159.1	0.135	0.893
CFMTp ✻ Aperture	CFMTp ✻ Aperture	-4.16e−4	8.91E-05	-5.91e−4	-2.42e−4	9218	-4.675	< .001
Phase1 ✻ Aperture	Recognition - Learning ✻ Aperture	0.01039	0.00253	0.00543	0.01535	8088.7	4.104	< .001
CFMTp ✻ Phase1 ✻ Aperture	CFMTp ✻ Recognition - Learning ✻ Aperture	-1.85e−4	1.78E-04	-5.34e−4	1.64E-04	9161	-1.041	0.298

Table C.19. Linear mixed model results for PC5 loading.

For the model of PC1, we set participants' intercept as a random effect and CFMT+, phase, and aperture size as fixed effects. Linear mixed models reveal no significant main effect of CFMT+ scores [b= 0.02, CI= [-0.18, 0.23], t(43.0)= 0.2, p= 0.844], no significant main effect of phase [b= 0.01, CI= [-0.023, 0.043], t(8521.1)= 0.6, p= 0.545], but a significant main effect of aperture size [b= 0.11, CI= [0.08, 0.134], t(653)= 7.75, p < 0.001]. These results show that PC1 strongly codes for the fixation pattern differences across aperture sizes. However, we found a significant the two-way interaction of CFMT+ and phase [b= -0.057, CI= [-0.086, -0.027], t(9155)= -3.77, p < 0.001]. As with Experiment 3a, the simple effects (using phase as a moderator) show a larger difference during face learning [b=0.05] than

recognition [b=-0.01], showing that higher CFMT+ scores possess higher PC1 scores during the learning phase of the experiment. The three-way interaction of CFMT+, phase and aperture size [b = 0.03, CI= [0.00, 0.06], t(9157.4)= 2.17, p= 0.030] was non-significant.

Gaze Dispersal Analysis (all apertures)

We investigated whether the amount of exploration differed across the face recognition ability spectrum. For that, we calculated the Gini coefficient (Lorenz, 1905) for each trial-level heatmap. As a reminder, lower Gini coefficients would indicate higher dispersion of fixations (i.e. higher exploration). Using linear mixed models, we found a significant main effect of CFMT+ scores [b= -.008, CI= [-.014, -.002], t(43.0)= -2.44, p=0.019]. This result means that those who showed higher exploration (lower Gini coefficients) also showed higher performance on the CMFT+. We also found a significant two-way interaction between CFMT+ and phase [b= .007, CI= [.005, .008], t(9141.6)= 10.44, p< .001]. The simple effects showed an effect of CFMT+ during the learning (b=0.011) but not during the recognition phase (b= 0.004).

PCA – Learning and Recognition (Only NV)

We analysed PC loadings using a Linear Mixed Model. For this analysis, we used the function: Loading ~ 1 + Phase + CFMTp + Phase:CFMTp + (1| ID). See Tables C.20 – C.24.

Linear Mixed Model results for PC1

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	3.43E-04	0.11046	-0.21616	0.21685	43.2	0.0031	0.998
Phase1	Recognition - Learning	0.00625	0.03619	-0.06469	0.07718	1562	0.17265	0.863
CFMTp	CFMTp	0.00556	0.00856	-0.01121	0.02234	43.2	0.64985	0.519
Phase1 ✻ CFMTp	Recognition - Learning ✻ CFMTp	-0.00334	0.00281	-0.00885	0.00217	1562	-1.18819	0.235

Table C.20. Linear mixed model results for PC1 loading.

Linear Mixed Model results for PC2

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	0.00141	0.07654	-0.14859	0.1514	43.6	0.0185	0.985
Phase1	Recognition - Learning	0.00238	0.04551	-0.08682	0.0916	1562.1	0.0523	0.958
CFMTp	CFMTp	-0.00106	0.00593	-0.01268	0.0106	43.6	-0.1784	0.859
Phase1 ✻ CFMTp	Recognition - Learning ✻ CFMTp	0.00567	0.00354	-0.00126	0.0126	1562.1	1.6024	0.109

Table C.21. Linear mixed model results for PC2 loading.

Linear Mixed Model results for PC3

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	4.27E-04	0.07239	-0.1415	0.14232	43.7	0.00589	0.995
Phase1	Recognition - Learning	-0.0014	0.04601	-0.0916	0.08878	1562.1	-0.03044	0.976
CFMTp	CFMTp	-0.00745	0.00561	-0.0184	0.00355	43.7	-1.32788	0.191
Phase1 ✻ CFMTp	Recognition - Learning ✻ CFMTp	-0.00602	0.00358	-0.013	9.89E-04	1562.1	-1.68325	0.093

Table C.22. Linear mixed model results for PC3 loading.

Linear Mixed Model results for PC4

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	-2.72e−4	0.03428	-0.06746	0.06692	46.7	-0.00795	0.994
Phase1	Recognition - Learning	5.73E-04	0.05127	-0.09992	0.10107	1562.2	0.01117	0.991
CFMTp	CFMTp	0.00397	0.00266	-0.00125	0.00918	47	1.49103	0.143
Phase1 ✻ CFMTp	Recognition - Learning ✻ CFMTp	-0.00371	0.00398	-0.01151	0.0041	1562.3	-0.93013	0.352

Table C.23. Linear mixed model results for PC4 loading.

Linear Mixed Model results for PC5

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	-3.02e−4	0.04518	-0.08885	0.08824	45	-0.00669	0.995
Phase1	Recognition - Learning	-0.00167	0.05032	-0.1003	0.09696	1562.1	-0.03315	0.974
CFMTp	CFMTp	-8.83e−4	0.0035	-0.00775	0.00598	45.1	-0.25199	0.802
Phase1 ✻ CFMTp	Recognition - Learning ✻ CFMTp	0.00228	0.00391	-0.00539	0.00994	1562.1	0.58181	0.561

Table C.24. Linear mixed model results for PC5 loading.

Experiment 4

AUC Analysis

In Experiment 4, we investigated the accuracy of 9 DCNNs performing the stimuli for different participant groups and six aperture size conditions. We show the linear mixed model in Table C.25 and its simple effects in Table C.26.

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	0.3314	0.0222	0.28796	0.3749	16.8	14.946	< .001
Source1	RDM - CT	-0.0523	0.0123	-0.07631	-0.0283	147	-4.266	< .001
Source2	SR - CT	0.0273	0.0123	0.00328	0.0513	147	2.227	0.027
Aperture	Aperture	0.0151	5.44E-04	0.01404	0.0162	147	27.765	< .001
Aperture²	Aperture²	-9.15e−5	4.64E-06	-1.01e−4	-8.24e−5	147	-19.715	< .001
Source1 ✻ Aperture²	RDM - CT ✻ Aperture²	3.92E-06	2.74E-06	-1.44e−6	9.29E-06	147	1.433	0.154
Source2 ✻ Aperture²	SR - CT ✻ Aperture²	-1.60e−6	2.74E-06	-6.96e−6	3.77E-06	147	-0.583	0.561

Table C.25. Linear mixed model results for AUC.

Simple effects of Source : Parameter estimates

Moderator levels				95% Confidence Interval				
Aperture	contrast	Estimate	SE	Lower	Upper	df	t	p
Mean-1-SD	RDM - CT	-0.051	0.01168	-0.07407	-0.02792	147	-4.37	< .001
	SR - CT	0.0268	0.01168	0.0037	0.04985	147	2.29	0.023
Mean	RDM - CT	-0.0437	0.0094	-0.06232	-0.02518	147	-4.66	< .001
	SR - CT	0.0238	0.0094	0.00525	0.04239	147	2.54	0.012
Mean+1-SD	RDM - CT	-0.0301	0.01169	-0.05322	-0.007	147	-2.58	0.011
	SR - CT	0.0183	0.01169	-0.00483	0.04138	147	1.56	0.12

Table C.26. Simple effects for AUC. We set aperture size as moderator for this analysis.

Information available Analysis

In Experiment 4, we investigated the information available for each stimuli image. We show the linear mixed model in Table C.27 and its simple effects in Table C.28.

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		β	df	t	p
				Lower	Upper				
(Intercept)	(Intercept)	12.3895	0.20372	11.9902	12.78879	0	44209	60.82	< .001
contAperture	contAperture	2.50757	0.00844	2.49102	2.52411	1.0624	44209	297.08	< .001
groupLabel1	SR - CT	3.28991	0.22131	2.85614	3.72368	0.1091	44209	14.87	< .001
groupLabel2	RDM - CT	10.96792	0.19778	10.58027	11.35558	0.3174	44209	55.45	< .001
contAperture²	contAperture²	-0.01645	7.25E-05	-0.01659	-0.01631	-0.5052	44209	-226.79	< .001
groupLabel1 ✻ contAperture²	SR - CT ✻ contAperture²	-3.88e−4	5.06E-05	-4.87e−4	-2.89e−4	-0.0241	44209	-7.66	< .001
groupLabel2 ✻ contAperture²	RDM - CT ✻ contAperture²	-0.00141	4.52E-05	-0.0015	-0.00132	-0.0466	44209	-31.22	< .001

Table C.27. Linear mixed model results for Information Available.

Simple effects of groupLabel : Parameter estimates

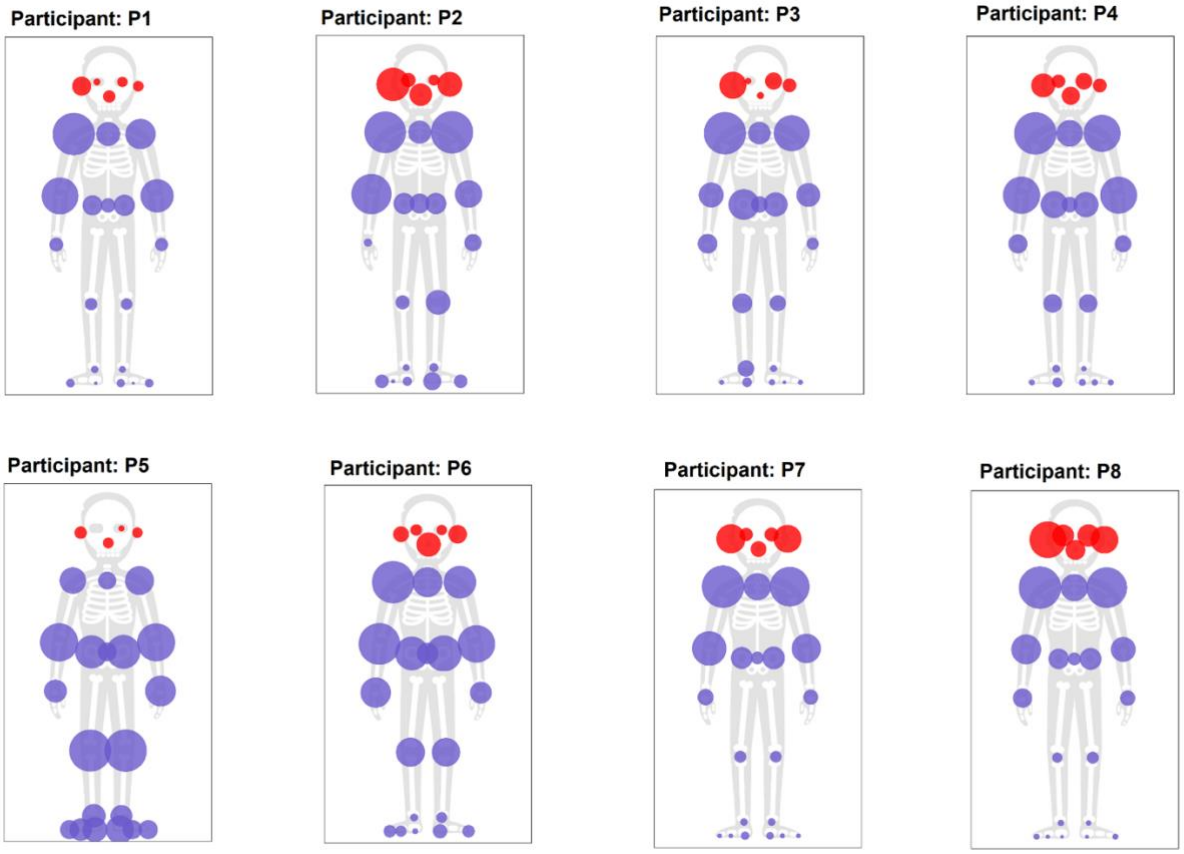
Moderator levels		Estimate	SE	95% Confidence Interval		β	df	t	p
contAperture	contrast			Lower	Upper				
Mean-1-SD	SR - CT	3.16	0.21	2.746	3.57	0.1249	44209	15.02	< .001
	RDM - CT	10.49	0.188	10.119	10.86	0.4149	44209	55.83	< .001
Mean	SR - CT	2.46	0.169	2.126	2.79	0.0973	44209	14.52	< .001
	RDM - CT	7.94	0.151	7.646	8.24	0.3143	44209	52.51	< .001
Mean+1-SD	SR - CT	1.16	0.21	0.745	1.57	0.0458	44209	5.51	< .001
	RDM - CT	3.21	0.188	2.843	3.58	0.127	44209	17.12	< .001

Table C.28. Simple effects for Information Available. We set aperture size as moderator for this analysis.

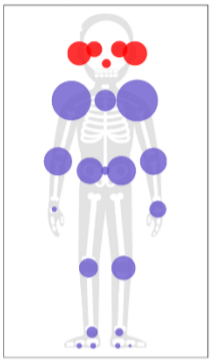
APPENDIX D

Individual visualisation of participants' body maps during the navigation task

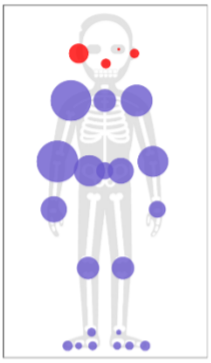
We observed high variability in the strategies engaged by participants when looking at people when navigating in the wild. Figure D.1 shows the proportion of fixations registered to each landmark separately for each participant in the navigation task.



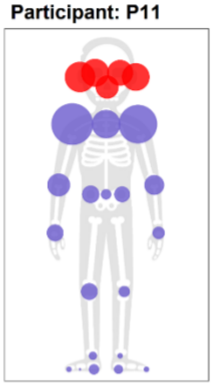
Participant: P9



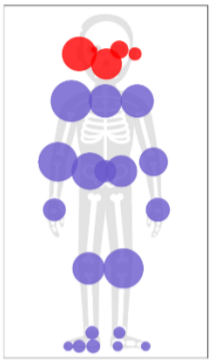
Participant: P10



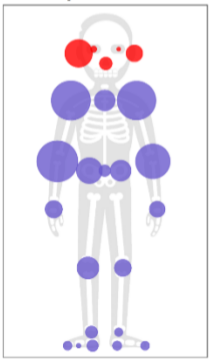
Participant: P11



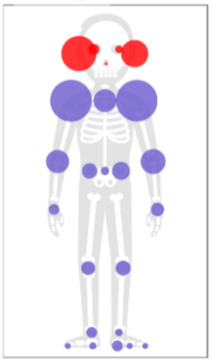
Participant: P12



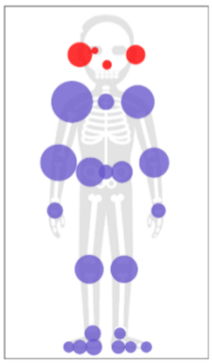
Participant: P13



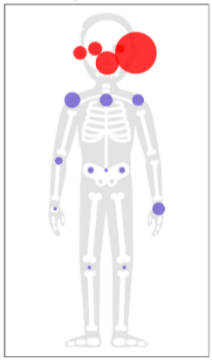
Participant: P14



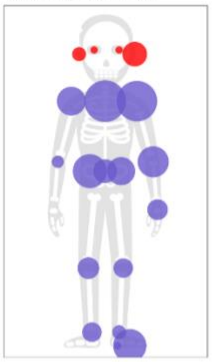
Participant: P15



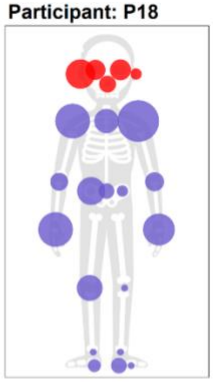
Participant: P16



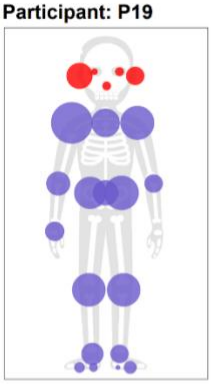
Participant: P17



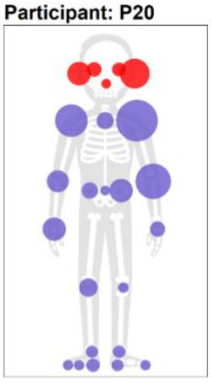
Participant: P18



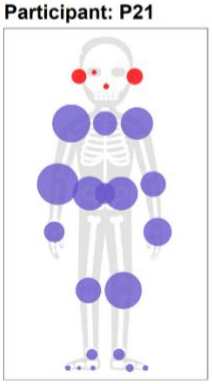
Participant: P19



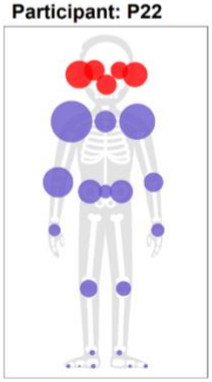
Participant: P20



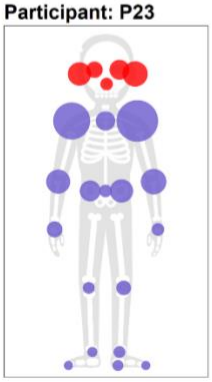
Participant: P21



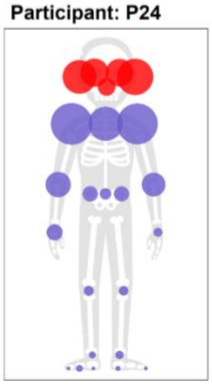
Participant: P22



Participant: P23



Participant: P24



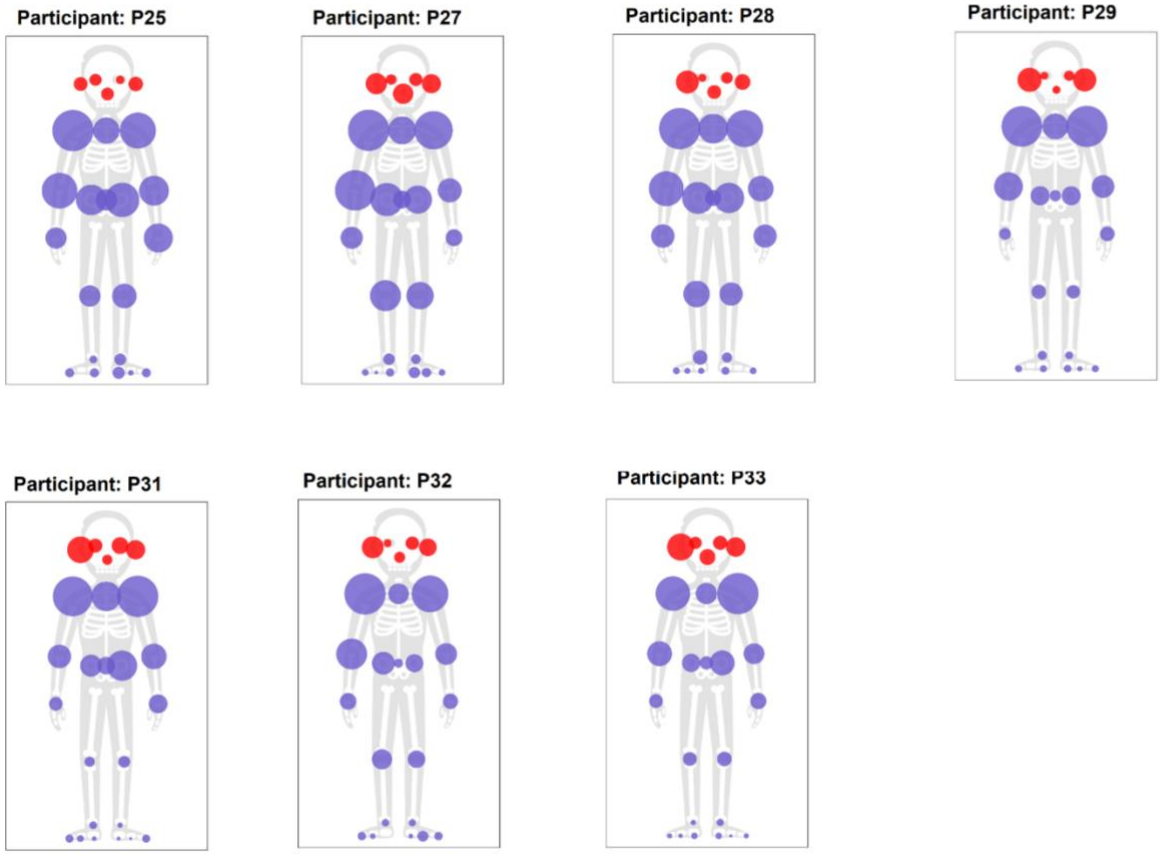


Figure D.1. Individual participant data during navigation task. For each participant, we show fixations on 25 dROI when viewing people during the navigation task. Face fixations are marked in red and body fixations in blue. The circle size for each dROI indicates the number of fixations participants made to that location.

Extended ANOVA analysis for 'Faces of passersby do not capture attention in live natural settings'

The main manuscript reports that participants were more likely to fixate on people in the navigation task when their faces were in full view. Still, we found no evidence that faces captured this attention more than other body regions.

This conclusion was supported by an ANOVA analysis of the data shown in Figure 5.2 comparing the probability of fixating on heads and bodies when faces were fully visible in a video frame versus when only partially visible due to head rotation or other occlusions. A 2 (Face type: part face, full face detected) X 3 (Fixation type: Head, Body, Not Person fixations) ANOVA revealed a significant interaction between face and fixation type, $F(2,60) = 9.76$, $p < 0.001$, $\eta^2_p=0.246$. Analysis of simple main effects showed a significant reduction of non-person fixations, $F(1,30) = 12.86$, $p < 0.001$, $\eta^2_p=0.300$, and an increase in both head and body fixations [Head: $F(1,30) = 7.035$, $p = 0.013$, $\eta^2_p=0.190$; Body: $F(1,30) = 6.64$, $p < 0.015$, $\eta^2_p=0.181$].

Extended analysis for 'Individual differences in naturalistic social attention' (analysis of residuals)

Figure D.2 shows scatterplots illustrating the individual differences analysis of residuals reported in the main text. In Figure D.2A, we calculated the linear regression model predicting the probability of fixating people (head and body) as a function of the average number of people detected in video frames, separately for the two route segments (First and Second). This analysis allowed us to calculate a residual value for each participant for each route segment. Figure D.2B shows the correlation between these residuals for the two route segments (Spearman's rho(29)= 0.532, p= 0.002), indicating that some participants tend to fixate more on people than others, regardless of the number of people they encountered on the walk.

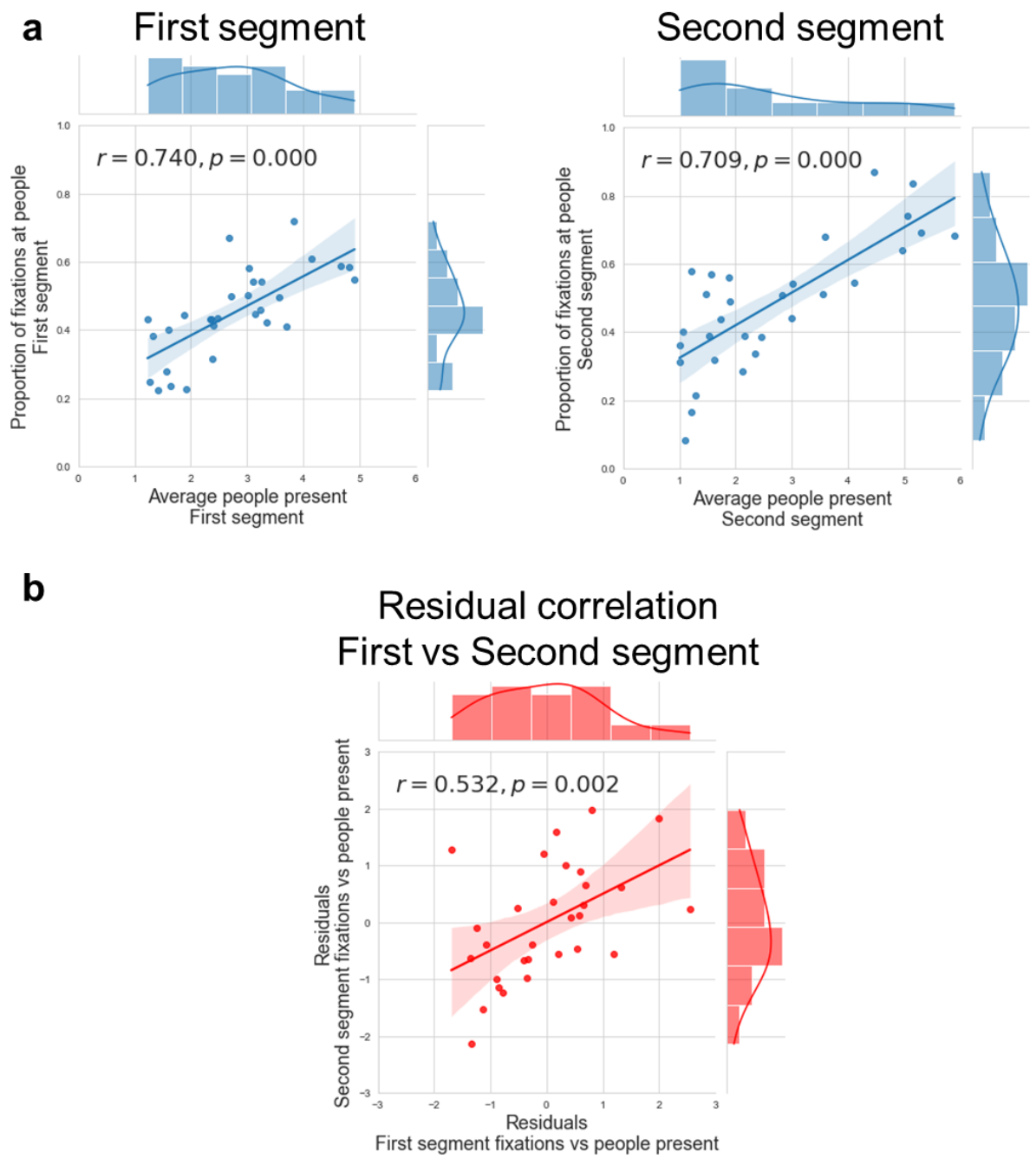
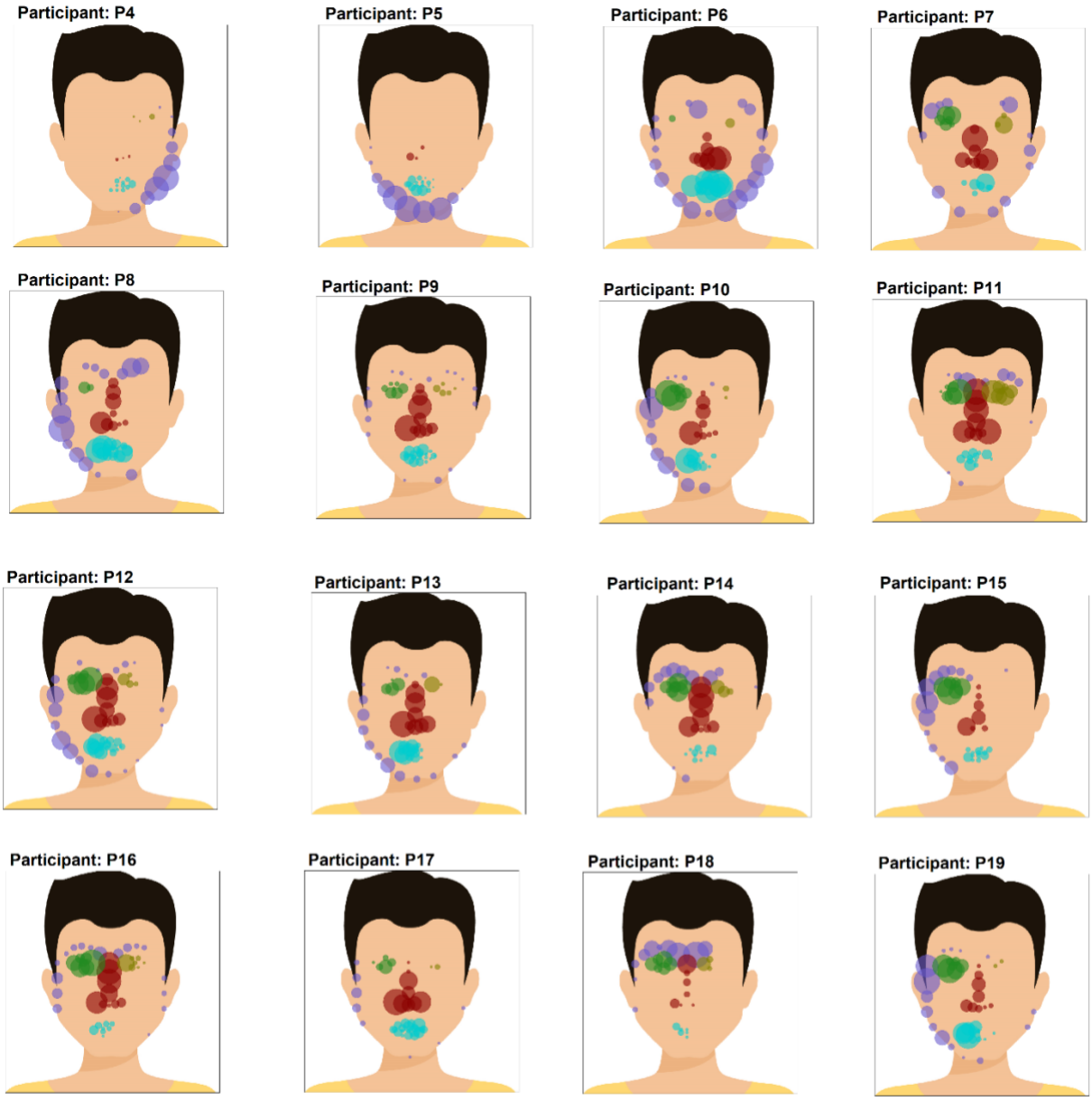


Figure D.2. Individual differences analysis of residuals. Panel A shows the proportion of fixations to people as a function of the average number of people present for each route segment. Panel B shows the correlation between the residuals found in panel A.

Individual visualisation of participants' facial maps during the face-to-face interaction task

We observed high variability in the fixation patterns shown by participants when engaging in a conversation with the experimenter. We filtered participants' recordings to analyse only fixation frames that contained the experimenter's face looking straight at the participant by using only frames where nose landmarks were detected by OpenPose (Cao et al., 2019). The

landmark registration method (see the main manuscript) detected 70 possible dynamic regions of interest (dROI) participants attended. Figure D.3 shows individual participant gaze patterns registered to facial landmarks.



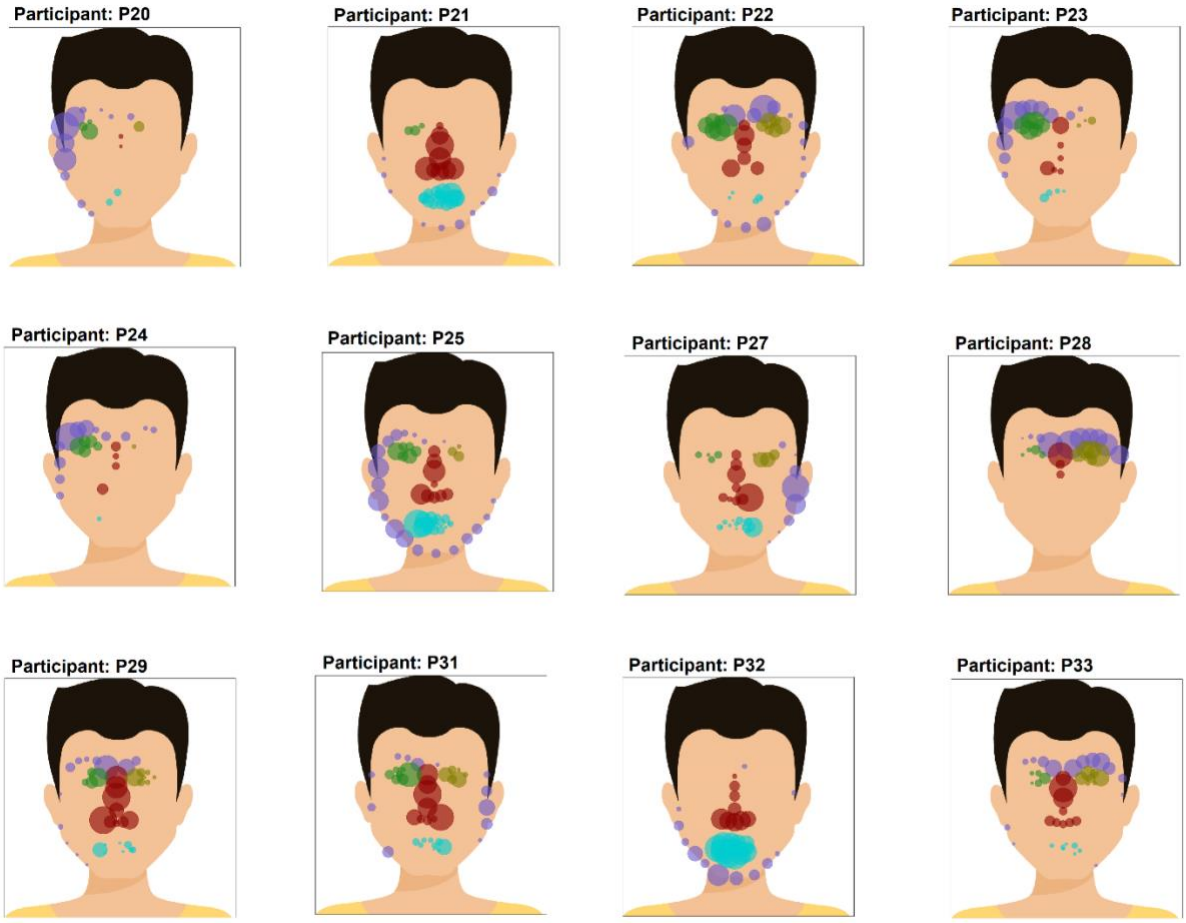
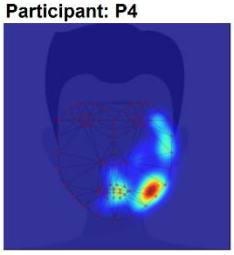


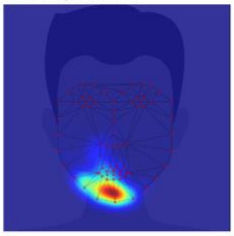
Figure D.3. Individual participant data during face-to-face interaction task. We show fixations to the 70 dROI identified by OpenPose while participants attended to the experimenter during the face-to-face interaction task. The colours delimit facial regions for this visualisation, and the size of the circles at each dROI indicate the number of fixations at that point.

The landmark registration in Figure D.3 is constrained to record 70 possible dROI positions on a face. However, the heatmap registration method allows more fine-grained analysis because it uses the relation between these landmarks to determine the exact location of where a fixation landed on a face (see main text). Figure D.4 shows individual heatmaps for each participant as they focused on the experimenter's face during the face-to-face task. Comparing the patterns in Figures D.3 and D.4, there is some indication of the advantage of using triangulation of spatial location. For example, when analysed using landmark registration, P32 appears to have a relatively diffuse gaze pattern, but this is revealed as a more focal pattern when using triangulation.

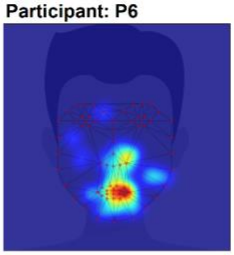
Participant: P4



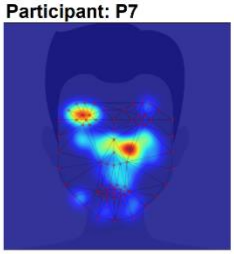
Participant: P5



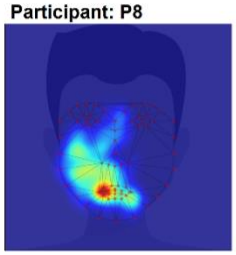
Participant: P6



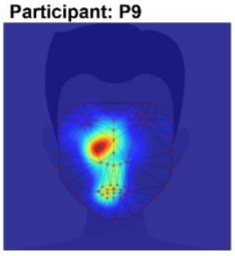
Participant: P7



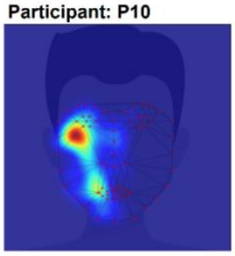
Participant: P8



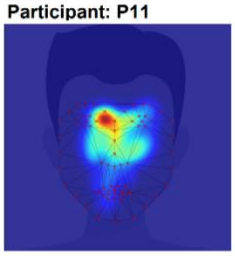
Participant: P9



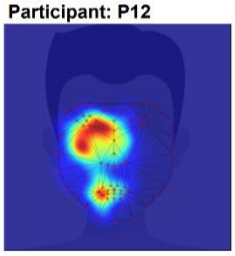
Participant: P10



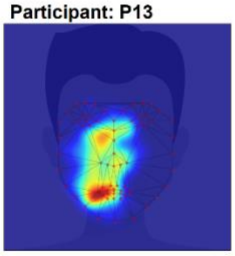
Participant: P11



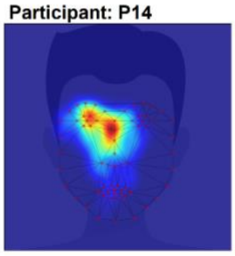
Participant: P12



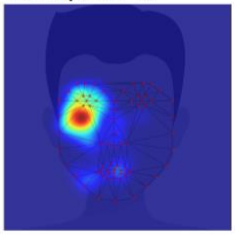
Participant: P13



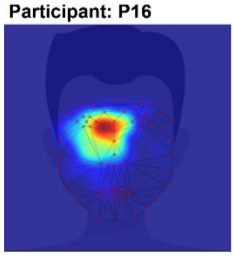
Participant: P14



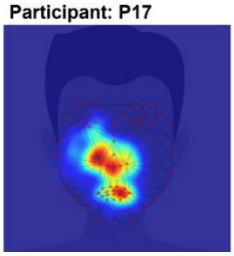
Participant: P15



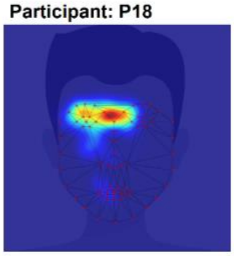
Participant: P16



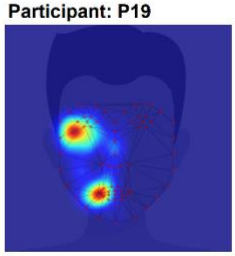
Participant: P17



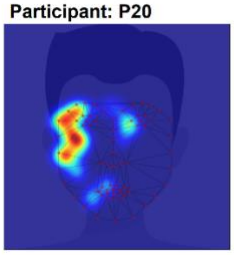
Participant: P18



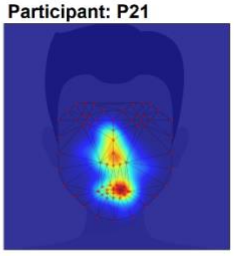
Participant: P19



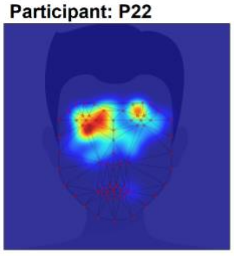
Participant: P20



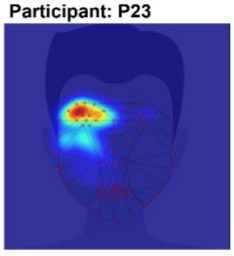
Participant: P21



Participant: P22



Participant: P23



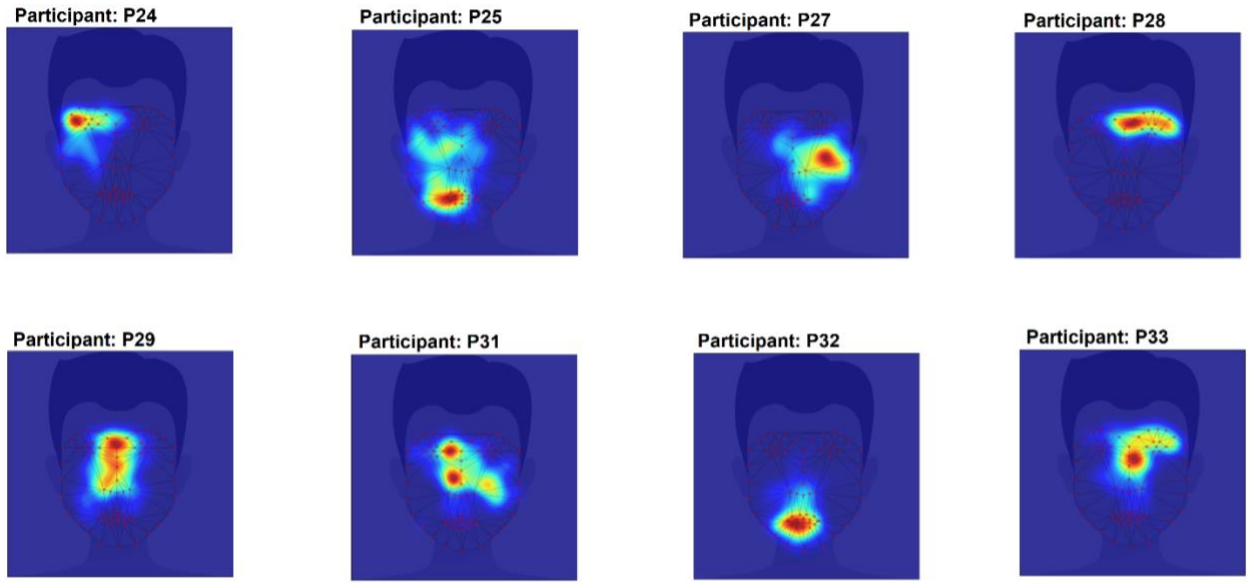


Figure D.4. Individual participant data during face-to-face interaction task. We show the heatmap registration method results to where participants attended when viewing a face during the face-to-face interaction task.

Comparing automatic versus manual coding

To validate the use of OpenPose to detect the presence of a person in a video frame, we randomly sampled 560 frames from participants' navigation task recordings in which a fixation was recorded. Figure D.5 shows an example of a frame image. Four naïve volunteers then manually count the number of people in every 560 frames. Figure D.6 shows the significant positive correlation between the average manual coding values with the automatic values provided by OpenPose ($r(559)= 0.89$, $p<0.001$).

Figure D.5. Example of a randomly selected video frame used to validate the OpenPose system.

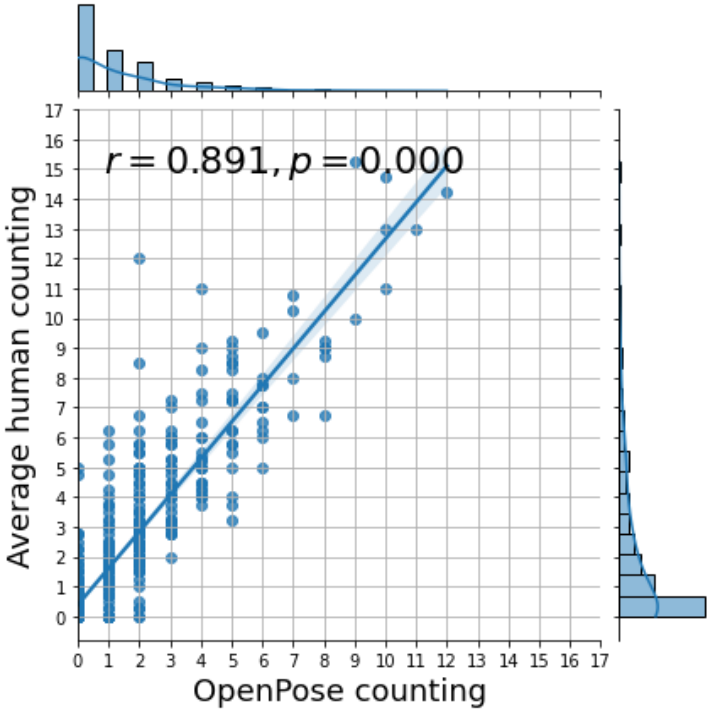


Figure D.6. Correlation between human vs algorithm in the number of people in the scene

References

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 172-186.