

Reconstructing face representations: a psychological and computational modelling approach to explore the effect of familiarity on human face recognition

By

Rosyl Selena Somai

University of Stirling

Psychology

Doctor of Philosophy

January 2023

Thesis abstract

When thinking about finding the face of a friend in a crowd, albeit challenging, most of us would be relatively certain we would be successful in finding them. However, if the same task were applied to an unfamiliar face, e.g. based on a photo, this would seem like an impossible task. Although the field of face recognition is very active across multiple disciplines, and there is a wealth of research on face recognition and the difference between familiar and unfamiliar faces, it still remains unclear what internal face representations we use. The aim of this thesis is to investigate what internal representations of faces we have and how these are different between familiar and unfamiliar faces. A better understanding of this internal representation will allow for the development of tasks that measure face processing abilities in more detail and increase our understanding of the visual information needed to identify a face, either familiar or unfamiliar.

In the behavioural part of this thesis, I focused on exploring the relationship between image-based and perceptual similarity. In the computational part, I explore the concept of familiarity in deep neural networks. This thesis can be summed up in four major findings: there is a linear relationship between image-based spaces and perceived similarity that can be used to explore the effect of face image transformations; familiar faces are perceived as more similar overall compared to unfamiliar faces; the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces; and the effect of familiarity can be found at the representational level (in DNNs). Although many questions remain open, it should be evident from this thesis that a computational cognitive approach to face recognition, although technically challenging and on many occasions laborious, will progress our understanding of human face recognition.

Acknowledgements

Thank you, Peter, for all your support. You have made my PhD journey possible by funding it through the Dylis Crabtree Scholarship, a scholarship you have used personal funds (your mother's heritage) to help women in science. I feel extremely fortunate to have been the recipient of this Scholarship. Thank you for always taking the time to explain the bigger pictures and go beyond just the PhD, the words of wisdom, the pictures of birds, your understanding and patience. It has been a challenging and highly educational process, and I cannot express my gratitude enough.

Thank you, Kevin, for taking on a Psychology PhD student and guiding me through learning challenging topics. It really helped me see new and exciting fields of science that I never expected to find myself in.

Thank you to everyone in the Face Lab. Coming from a different country can have some challenges, but luckily there was always someone around willing to help. In particular, thank you Anna, Ailsa, and Vicky for the warm welcome in the group and always checking in. Thanks, Ant, for always making time to talk stats when I once more confused myself! Thank you, Sandy, for all the chats, words of wisdom, and yoga exercises! I could not have wished for a better PhD buddy. Thanks, Jud, for showing how much I have learned and always being there with a helping hand. Thank you to all the amazing students who were part of the lab during these last years, you have kept me motivated to learn and I look forward to hearing about all your amazing future achievements. All of the support made this journey much more enjoyable, so thank you all.

Last but not least, thank you to my family and friends – here and in the Netherlands – who kept me sane during this progress. I feel privileged to say that I will refrain from using names because the list would be too long. But thank you. Especially during the times of lockdown, I highly appreciate all the support, words and gifts of encouragement, shared suffering, and memes. Thank you.

Summary of the dissemination

What follows is a summary of the dissemination of the research presented in this thesis.

Publication

Somai, R. S., & Hancock, P. J. (2022). Exploring perceptual similarity and its relation to image-based spaces: an effect of familiarity. *Visual Cognition*, 30(7), 443-456.
<https://doi.org/10.1080/13506285.2022.2089416>

Conference presentations

Somai, R. S. (March, 2023). Invited poster presentation “*Exploring sensitivity to change and its relation to image-based spaces*”, Workshop on Humans, Deep Networks & Face Recognition, London, UK.

Somai, R. S. (December, 2021). Poster presentation “*Familiarity and quantifying sensitivity to change*”, NVP the Dutch Society for Brain and Cognition conference, Egmond aan Zee, The Netherlands.

Somai, R. S. (August, 2021). Invited talk “*Exploring perceptual similarity and its relation to image-based spaces: an effect of familiarity*”, Applied Face Cognition Lab Talk series, virtual.

Somai, R. S. (May, 2021). Poster presentation “*Exploring perceived face similarity and its relation to image-based spaces: an effect of familiarity*”, Vision Sciences Society meeting, virtual.

Somai, R. S. (December, 2019). Poster presentation “*Beyond binary face recognition task: The effects of familiarity on sensitivity to subtle face changes*”, NVP the Dutch Society for Brain and Cognition conference 2021, Egmond aan Zee, The Netherlands.

Somai, R. S. (December, 2019). Poster presentation “*Beyond binary face recognition task: The effects of familiarity on accuracy of face memory*”, European Conference of Visual Perception, Leuven, Belgium.

Table of contents

Contents

Thesis abstract.....	3
Acknowledgements.....	4
Summary of the dissemination.....	5
Table of contents.....	6
List of tables.....	10
List of Figures	12
Part I.....	14
1. Chapter 1 - Introduction - the psychology of face representations.....	15
1.1. Introduction	15
1.2 Are faces special?.....	16
1.3. What is familiarity?.....	18
1.4. Differences and similarities between familiar and unfamiliar face perception.....	20
1.5. Facial variation.....	25
1.6. Two ends of a scale – measuring individual differences, deficits and enhanced performance.....	29
1.7. Exploring face representations through (psychological) models	36
1.8. Structure and context of this thesis	42
2. Chapter 2 - Beyond binary face recognition task: The effects of familiarity on sensitivity to subtle face changes	46
Chapter main message.....	46
2.1. Introduction	47
2.2. Methods.....	50
2.3. Results	55
2.4. Discussion Experiment 1.....	57
2.5. Experiment 2	58
2.6. Methods experiment 2.....	65
2.7. Results experiment 2	66
2.8. General discussion.....	67
3. Chapter 3 - Exploring perceptual similarity and its relation to image-based spaces: an effect of familiarity.	69
Chapter main message.....	69

3.1. Introduction	70
3.2. Experiment 3 - exploring the effects of single component transformations	73
3.3. Methods.....	73
3.4. Results Experiment 3	77
3.5. Discussion Experiment 3.....	81
3.6. Experiment 4a: the relationship between the number of principal components transformed, perceptual similarity and familiarity	82
3.7. Methods Experiment 4a	84
3.8. Results and discussion Experiment 4a	87
3.9. General discussion.....	92
4. Chapter 4 - Exploring other aspects of perceptibility of change: magnitude of change	95
Chapter main message.....	95
4.1. Introduction	95
4.2. Methods Experiment 4b	97
4.3. Results Experiment 4b	101
4.4. Discussion Experiment 4b.....	102
5. Chapter 5 - Title: Familiarity and quantifying sensitivity to dissimilarity	104
Chapter main message.....	104
5.1. Introduction	105
5.2. Methods Experiment 5a: sensitivity to change within identity (2 SD change).....	106
5.3. Results Experiment 5a.....	109
5.4. Discussion Experiment 5a.....	110
5.5. Experiment 5b: crossing the identity boundaries (4 SD change)	111
5.6. Results Experiment 5b	112
5.7. Discussion Experiment 5b.....	113
5.8. General discussion.....	114
6. Chapter 6 - Finding the category boundaries of faces: familiarity and the distinctiveness of transformations	117
Chapter main message.....	117
6.1. Introduction	118
6.2. Methods Experiment 6a	122
6.3. Results Experiment 6a.....	125
6.4. Discussion Experiment 6a.....	129

6.5. Methods Experiment 6b	130
6.6. Results Experiment 6b	131
6.7. General discussion.....	134
Part II	137
7. Chapter 7 - Review on the use of deep neural networks in understanding human face recognition	138
Chapter main message.....	138
7.1. Introduction	138
7.2. Stages of artificial face recognition.....	139
7.3 Why DNNs?	140
7.4. Why now?	143
7.5. Current directions	145
7.6. Summary and conclusions.....	152
8. Chapter 8 - Exploring the representation of familiar faces using deep neural networks.....	153
Chapter main message.....	153
8.1. Introduction	153
8.2. Methods Experiment 7a	156
8.3. Results Experiment 7a.....	162
8.4. Discussion Experiment 7a.....	162
8.5 Experiment 7b	163
8.6. Methods Experiment 7b	163
8.7. Results Experiment 7b	164
8.8. Discussion Experiment 7b.....	165
8.9. Experiment 7c	166
8.10. Methods Experiment 7c	166
8.11. Results Experiment 7c.....	167
8.12. General discussion and evaluation of this framework	167
9. Chapter 9 - The relationship between distance measures of representational space and behavioural data	174
Chapter main message.....	174
9.1. Introduction	174
9.2. Methods.....	180
9.3. Results	182

9.4. Discussion	187
10. Chapter 10 - Conclusions and future directions	189
10.1. Summary of chapters.....	189
10.2. Difference between familiar and unfamiliar faces	191
10.3. Future directions in computational face recognition research	192
10.4. Brief summary of the recent changes in the field face identity recognition	202
10.5. Conclusions	204
Supplementary materials.....	205
Supplementary material – Chapter 3.....	205
Supplementary materials – Chapter 4	207
Supplementary materials – Chapter 8	209
Supplementary materials – Chapter 9	216
Supplementary materials - General	218
Glossary	223
References	224

List of tables

Table 1. Average similarity rating per Transformation Level.	101
Table 2 components per pair distance.....	109
Table 3. Averages per number of components changed.	112
Table 4. averages per pair and Familiarity.....	114
Table 5. average similarity rating per components changed.....	115
Table 6 Familiarity rating per ID	126
Table 7 Familiarity rating per ID	133
Table 8 Distribution of familiarity in experiment 4a.	205
Table 9 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 4a.	205
Table 10 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 4a.	206
Table 11 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 4a.	206
Table 12 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 4b.	207
Table 13 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 4b.	207
Table 14 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 4b.	208
Table 15 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 7a.	209
Table 16 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 7a.	209
Table 17 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 7a.	210
Table 18 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level in Euclidean distance with Transformation Level 2 as reference for Experiment 7b.....	211
Table 19 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 7c.	211

Table 20 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 7c.	212
Table 21 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 7 as reference for Experiment 7c.	212
Table 22 Performance of the models in Experiment 7a.....	213
Table 23 Performance of the models in Experiment 7b	214
Table 24 Performance of the models in Experiment 7c.....	215
Table 25 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in cosine similarity, with Transformation Level 2 as reference for Experiment 7b. ..	216
Table 26 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Manhattan distance, with Transformation Level 2 as reference for Experiment 7b.	216
Table 27 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Jaccard-Global distance, with Transformation Level 2 as reference for Experiment 7b.....	217
Table 28 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Jaccard-Detail distance, with Transformation Level 2 as reference for Experiment 7b.....	217
Table 29 Overview of findings	218

List of Figures

Figure 1 Example of the subtle changes created.....	52
Figure 2 Example of 1 ‘step’ from the original.	53
Figure 3 Overview of one trial.....	54
Figure 4 Violin plots of difference score per type and familiarity level.....	56
Figure 5 Overview of one trial.....	59
Figure 6 Violin plots of difference score (left) and the reaction time (right) per type and familiarity level in the pilot.....	60
Figure 7 Violin plots	61
Figure 8 Overview of a trial in the second pilot	62
Figure 9 Overview of a trial in the final pilot and subsequent experiment.....	62
Figure 10 Violin plots	63
Figure 11 The raw scores in the high PS condition of the final pilot split in familiar (top) and unfamiliar (bottom) per identity.....	64
Figure 12 Violin plots of difference score (left) and the reaction time (right) per type and Familiarity level in experiment 2. The white dot indicates the median.....	66
Figure 13 Example of the transformations used.	75
Figure 14. Example of a trial	76
Figure 15 Example of cropped images, ...	77
Figure 16 Correlation between perceptual similarity and Gabor distance.....	79
Figure 17 Correlation between perceptual similarity and Pixel distance.	80
Figure 18 Correlation between perceptual similarity and DNN similarity.....	81
Figure 19 Gabor distance between the original and altered images	85
Figure 20 An example of the transformation.....	86
Figure 21 Overview of a trial.....	87
Figure 22 Correlation between perceptual similarity and Gabor distance. The trendline is fitted to the data of all change size groups.	88
Figure 23 Correlation between perceptual similarity and Pixel distance.	89
Figure 24 Correlation between DNN distance and perceptual similarity.....	90
Figure 25 Average similarity rating per Transformation Level grouped by familiarity.....	90

Figure 26 Schematic visualization of the image transformation for the stimuli creation of Experiment 4a (left) and Experiment 4b (right, after correcting for extreme values).	99
Figure 27 Example of the PCA transformation using 2 SD and a change towards the average.	100
Figure 28 Overview of a trial.....	100
Figure 29 Average similarity rating per Transformation Level grouped by familiarity.....	101
Figure 30 Example of the transformations.....	107
Figure 31 Example of a trial	108
Figure 32. Proportion correct per pair distance.	109
Figure 33 Example of the transformations.....	112
Figure 34. Proportion correct per pair distance	113
Figure 35 Example of the transformations used.	123
Figure 36 Example of a trial in which 2 components are changed (2PCs condition).....	124
Figure 37 Example of a trial in which 5 components are changed (5PCs condition).....	125
Figure 38 Components selection (frequency) per ID, change size and Familiarity.....	127
Figure 39 Components selection (%) per ID, change size and Familiarity.	127
Figure 40 Gabor distances per ID, change size and component.	128
Figure 41 DNN similarity scores per ID, change size and component.....	129
Figure 42 Components selection (frequency) per ID, change size and familiarity.	132
Figure 43 Components selection (%) per ID, change size and familiarity.	132
Figure 44 Gabor distances per ID, change size and component.	133
Figure 45. Schematic example of a DNN.	141
Figure 46 Example of the transformations.....	158
Figure 47 Example of creating one pair of subsets.....	159
Figure 48 Euclidean distance per Transformation Level.....	162
Figure 49 Euclidean distance per Transformation Level.....	164
Figure 50 Euclidean distance per Transformation Level.....	167
Figure 51 Example of the rounding for the Jaccard computation.....	179
Figure 52 Recap of the findings of Experiment 4a, 5a and 5b.	184
Figure 53. Plots of the different distance measures per Transformation Level.	186

Part I

1. Chapter 1 - Introduction - the psychology of face representations

1.1. Introduction

I remember well when I went to school wearing my glasses for the very first time. I was absolutely dreading all the responses (I was in high school at the time). However, this did not happen, at all. Since it is not in the nature of Dutch people to refrain from responding out of politeness per se, I asked some people about it. And there it was: realization and surprise, they had not noticed the difference. To be fair, some noticed “something”, but very few people actually spotted the change because my face looked more natural/like me with glasses, was the most-heard comment. I am recounting this little anecdote as it captures a lot about the nature of face representations: unfamiliar people will see the difference and familiar people have a robust representation that will help them recognize you under a wide range of variation.

Although the field of face recognition is very active across multiple disciplines, and there is a wealth of research on face recognition and the difference between familiar and unfamiliar faces, it still remains unclear what internal face representations we use. The aim of this thesis is to investigate what internal representations of faces we have and how these are different between familiar and unfamiliar faces. In this review that constitutes Chapter 1, I will explore what is known about the internal representation of faces and how these are affected by familiarity. The overview starts with a fundamental and widely debated question “Are faces special?”, as the different answers to this question allude to the complex nature of face representations and provide a good overview of possible ways to explore face processing. After this, definitions of familiarity and known differences between familiar and unfamiliar face processing are discussed. This is followed up with a discussion on perceivable facial variation, what we can learn about face representation and familiarity from people with prosopagnosia and ‘super-recognizers’, and how face matching ability is measured. Lastly, the review is closed with a look at psychological models of face recognition that attempt to capture much of the knowledge discussed and a brief overview of the structure of this thesis.

1.2 Are faces special?

Section main message: Faces are special from a practical, social and societal perspective but it remains to be seen how special they are cognitively and what role familiarity plays in that.

“Are faces a special category for our brain?” is a question asked many times in the field of face perception. In daily life, we encounter many faces, of passers-by, friends, family and more, that we on many occasions need to recognize as an individual. This task might seem easy due to the ease with which we process faces, but this becomes quite astonishing if you focus on how similar faces are in a general sense. The necessity of a high-level within-category recognition at a subordinate level is described as what makes faces a special category indeed (Young & Burton, 2018a). Although one might be intuitively inclined to say humans are experts in recognizing faces because of this necessity, there is still a lively debate about whether or not this sophisticated skill is specific to faces, or if it generalizes to other objects such as cars and sheep (e.g. Mckone & Robbins, 2011; Rossion, 2018; Sunday & Gauthier, 2018; Young & Burton, 2018b, 2018c, 2018a)? Many different answers with a wide range in level of nuance have been proposed.

One study, that might be one of the most famous explorations of face expertise, is the work on ‘Greebles’ by Gauthier and Tarr (1997). In this study, fictitious creatures dubbed ‘Greebles’ were developed to study if the necessary within-category recognition at a subordinate level we use for faces also generalizes to stimuli that are not faces. Greebles had different levels of categories similar to faces: as a class they were Greebles that could be distinguished in family, gender and individual. Participants were trained in different conditions to become ‘Greeble’-experts and performance between these experts and novices were then compared. Their findings show that expertise led to a more accurate and faster response in a forced-choice recognition task. Additionally, they found greater sensitivity to configural changes in experts compared to novices. Thus, the authors argue that these effects are not face-specific, but rather linked to the characteristics of any visually homogenous category.

Even if you assume the point of view that the level of expertise we have for faces *could* be generalized to another category as long as it carries the same characteristics, it raises the question of whether it *would* be. In this thesis, the neural representation of face perception is only discussed in a limited way, but the findings in the discussed literature do show dedicated brain

areas that to this day appear to respond only to faces, such as the fusiform face area (Loffler et al., 2005). One theory argues that the brain operates on the free-energy principle (Friston, 2009, 2010; Sengupta et al., 2013), meaning it will try to minimize the cost of operations, and outcomes are limited to what is necessary for the operation. Thus, the presence of dedicated brain areas for face processing is an indication of the importance of faces, because the brain would not spend so many costly operations on anything that is not vital for our day-to-day life.

One of the major uses of visual facial information is social perception. Even in unfamiliar faces, people can observe expression (e.g. Vicki Bruce, 1986), trustworthiness, status (e.g. Palomares & Young, 2018) and attractiveness (e.g. Jenkins et al., 2011) and more. Insight on what visual information is used to make these judgements remains limited, though some interesting work has been done in studies on expression (e.g. the relevance of a V-shape in the perception of anger; Larson et al., 2007; Lobue et al., 2010). This purpose of faces directly influences the representation, as it clearly requires a complex visual representation to interpret social constructs. Thus, faces are most likely a special category, as there is as of yet no equivalent visually homogenous group with the same social importance.

Apart from the theoretical reasons in this discussion, a practical reason for the special status faces hold is of relevance to this project: faces are special because of the sheer amount of research on the topic by a multitude of disciplines. The interest in face recognition, human and artificial, has increased due to more daily life implementations. For example, it's common to have your smartphone unlock with face recognition, go through border control with automatic identity verification and an increasing number of countries use face recognition as a security tool. So, although the verdict is out on how special face processing exactly is a cognitive function in the brain, faces are without a doubt special due to societal relevance.

In sum, the debate on whether or not faces are special, and if we have expertise in specifically the processing of faces, is focusing on terminology, task design and attempts to divide and define mechanisms of face perception (e.g. Mckone & Robbins, 2011; Rossion, 2018; Sunday & Gauthier, 2018; Young & Burton, 2018b, 2018c, 2018a). This thesis does not set out to answer questions on expertise in face processing compared to other categories, but will inevitably contribute to the ongoing discussion, as a successful model will lead to insight in the mechanisms of face processing and can provide a framework for further exploration (such as the work of Blauch et al., 2021b) which I will discuss in further detail in Chapter 7. As touched upon

before, a key concept in the debate on the singularity of face processing is familiarity, as some argue that ‘face’ processing skills can generalize if you have an expert level of knowledge (familiarity) with another object. Before we dive deeper into differences between familiar and unfamiliar face processing, we will discuss the current definitions of familiarity.

1.3. What is familiarity?

Section main message: different operational definitions for ‘familiar’ are being used and it is important to keep this in mind when comparing studies.

On a basic level, this question “What is familiarity” is very straightforward answer, (face) familiarity means having good knowledge of a face due to experience. However, this definition is quite broad and ambiguous, because what does it mean to have ‘good knowledge’ of a face and what constitutes ‘experience’? This loose definition of familiarity becomes very clear when reviewing the literature on differences between familiar and unfamiliar face processing. In this section, I will discuss what definitions are used for familiarity and attempts within the field of face recognition to unify these definitions.

Familiarized

In several studies, the term ‘familiarized’ is used instead of ‘familiar’ and tends to refer to a novel face that goes through a learning phase and then is dubbed ‘familiarized’. At the end of the familiarization, there is a check to see if the participant actually remembers the face. For example, this can be done with a new/old task or by learning the face in combination with a name and asking for identification. Even within the definition of ‘familiarized’ there is a range, as some studies use a learning face that repeatedly presents the same single image (e.g. Abudarham & Yovel, 2018) and others argue that you need to present multiple instances of a face as the variability is what makes a face familiar (e.g. (Ritchie & Burton, 2017)). As we will discuss later, in 1.7 “Exploring face representations through (psychological) models”, the semantic and additional knowledge (such as episodic memory) is considered part of what makes

a face familiar too. Thus, one can wonder how ‘familiar’ a face is when the experience is based on a very limited exposure with the face in an experiment.

Famous faces

Famous faces are faces of people well-known in the public domain and are often shown through media, such as faces of actors and actresses, politicians and royal family. Although we might feel we would always recognize a certain person because we are familiar with them from the media, this might actually not be true (Carbon, 2008). Carbon (2008) used two experiments with personally familiar and famous faces to test recognition rates on uncommon photos and modified photos and found that the recognition rate dropped hugely compared to the personally familiar faces. However, the initial performance of the famous faces was a lot higher than the personally familiar faces. This has most likely to do with the faces used as ‘personally familiar’ as these were lecturers from the university (to keep image information the same for all participants) and the type of exposure the famous people had, which I will discuss in more detail in the subsection ‘A continuum of familiarity’.

Personally familiar: family and friends

The category considered most familiar are personally familiar people: people you have many interactions with in real life that causes you to be exposed to much variability in their face. Additionally, you have a personal connection to these people, thus they have some social importance to you, and you have additional knowledge beyond a visual representation of them.

A familiarity continuum

The order in which each of these definitions of familiarity are discussed is no coincidence, but it is based on the idea of familiarity on a continuum. This idea was already described by Rhodes in 1985 who stated “To the extent that faces vary on a familiarity continuum, more semantic information would be available for more familiar faces.” Later, Clutterbuck and Johnston (2002, 2004) experimentally showed that different levels of familiarity influence task performance and can be used as an index for familiarity. Thus, there appear to be different levels of familiarity, although it still remains unclear what each level of familiarity characterizes.

In the section above, I already touched upon the broad operational definition of ‘personally familiar’, as it appears even this category is not straightforward. As Carbon (2008) already suggests, the type of exposure matters. However, he argued that famous faces are less robust than personally familiar faces, but since 2008 when the article came out, how we consume media content has changed immensely. For example, where in the past people on TV were considered famous, now you also have famous YouTubers, Instagrammers and TikTokkers. These content creators often focus on building a more intimate connection with their “followers”, have little other than themselves on screen (some in more variable situations than others) and often share personal information. Comparing this to the attributes normally associated with ‘personally familiar’ we see a huge overlap, but are missing real life interactions and one-to-one interactions. However, with the current global situation, many of our personal interactions have moved to an online platform too. It remains unclear to what extent internal representations differ per exposure type. Therefore, the continuum of familiarity seems a useful approach but only in a limited sense as long as it remains unclear how faces go from novel to familiar. For the purpose of this PhD project, I use famous faces as familiar faces, specifically YouTubers in the first 2 experiments and actresses thereafter. In the next section, we discuss the differences between familiar and unfamiliar face perception in more depth.

1.4. Differences and similarities between familiar and unfamiliar face perception

Section main message: Although differences between familiar and unfamiliar face processing are evident, defining qualitatively what is different is difficult and has been researched to a limited degree.

It is now broadly accepted that there is a difference between familiar and unfamiliar face perception, although the extent of the differences remains unclear. Even from the differences that are found, not all agree to the extent how ‘poor’ unfamiliar face processing is (for example, see Rossion, 2018; Young & Burton, 2018b). For that reason, the overview below is limited to studies with designs that allow for direct comparison of familiar and unfamiliar faces for as much as this is possible.

Logically speaking, an unfamiliar face has no prior internal representation if the face is unseen before. Additionally, since we remember familiar faces, there must be some internal representation. Since a face is linked to an identity, more information is stored than just a representation of a face, we also hold semantic information. Thus, the difference between familiar and unfamiliar faces must lie in the internal visual representation we have of a face, but - quite obviously- also in the semantic representation we hold of an identity. This concept will be discussed to a greater extent in section 1.8 “Exploring face representations through models” and for the current discussion we limit ourselves to only research that explores the effect of (manipulated) visual information and its relationship with familiarity.

Differences in task performance

One of the most well-known differences related to familiarity from early research is that familiar faces are responded to faster than unfamiliar faces (Bruce, 1986). In a series of experiments, participants were tasked to make judgments of expression, sex, and facial configuration (intact configuration or jumbled). An effect of familiarity was found for the judgement of sex and facial configuration, but not for judgements of expression.

Later research attempted to find out what different visual information led to these differences, which has focused on internal versus external features. It has been found (and widely replicated) that recognition of a familiar face is higher when using internal features instead of external features and that this difference is not present for unfamiliar faces (for an overview, see (Osborne & Stevenage, 2008)). This has been dubbed the Internal (over external) Feature Advantage (IFA).

However, some evidence showing the contrary has also been found. Before studying the effects of familiarity, Abudarham and Yovel (2016) developed a framework that aimed to define what critical features are used for identification. First, they created a face space of 20 features. Then, similarity judgements were collected on different manipulations of these features, between and within faces. They found that the distances in their face space correlated with perceptual sensitivity, thus validating the choice for the 20 original features that went into their face space.

This method of perceptual similarity was then applied to explore the difference between familiar and unfamiliar face identification (Abudarham & Yovel, 2018). After a familiarization phase, participants were shown three different types of pairs of faces. The pairs were either two

different images of the same face, one original and one changed image of the same face, or two images each of a different face. They were asked to rate similarity using a 6-point scale that read as follows: “1 Definitely same person; 2 Same person; 3 Probably same person; 4 Probably different people; 5 Different people; 6 Definitely different people”. Images were presented on screen simultaneously. They found a high correlation between perceptual sensitivity and their face space distances. Additionally, perceptual similarity scores of familiarized faces correlated strongly with perceptual similarity scores of novel faces that were obtained in previous research (Abudarham & Yovel, 2016). However, it must be noted that familiar here meant a previously unfamiliar face that underwent a familiarization phase that exposed the participant repeatedly to the same single image of the identity in combination with a name. After the familiarization, participants were able to name the face 97% of the time. As discussed in section 1.3, this is a limited interpretation of ‘familiar’.

Apart from different definitions of familiarity in the studies mentioned in this section, there is a difference in how the visual information contained in faces is broken down into categories. The studies that show a difference related to familiarity categorized visual input in the categories ‘internal’ versus ‘external’ features, and the studies showing no difference used a categorization based on perceptual sensitivity (as defined by Abudarham & Yovel, 2016). Although it appears these results contradict each other, it could well be possible that the IFA is mediated by perceptual sensitivity and that the same pattern of response could be seen if mediating effects are taken into account. Thus, it appears the way visual information is categorized affects whether or not you are going to find an effect of familiarity. The authors (Abudarham & Yovel, 2018) point this out themselves by stating “These findings suggest that a gross categorization to internal and external features is not sensitive for revealing which features are used for face recognition.” However, the same argument might be made for the lack of finding differences between familiar and unfamiliar faces, because yet a different categorization of the visual information contained in faces could show an effect of familiarity. For example, in a study done on configurational information, an effect of familiarity was found (Brooks & Kemp, 2007).

In a 2-alternative forced-choice paradigm, Brooks and Kemp (2007) explored the effect of familiarity on sensitivity to displacement of features. This is an important study for this project, as it is the only study I am aware of on the effect of familiarity in relation to such a well-

defined and selective part of the visual information (feature displacement) contained in a face, that can also be controlled on a gradual scale. Brooks and Kemp (2007) found that sensitivity increased for displacement of the eyes and the nose in familiar faces compared to unfamiliar faces. This work is especially relevant for Experiment 1 and 2, so I will discuss it in further detail in Chapter 2.

Lastly, the low performance in face matching unfamiliar faces is also described to a great extent in literature on eyewitness testimony. These findings are also replicated and further explored in lab settings, where line-ups are often replaced with a one-to-many array in which the participant is asked to match 1 target to one of 10 pictures (e.g. Bruce et al., 1999, 2001). In this study, there was no direct comparison between familiar and unfamiliar faces but only unfamiliar faces were used, which makes it difficult to state anything about the differences between familiar and unfamiliar face matching. However, with a 30% error commonly found, it should be clear that face matching tasks with unfamiliar faces appear to be a lot more difficult than with familiar faces. Other lines of research show that accuracy in identifying/matching unfamiliar faces drops when there are changes in viewpoint, lighting, context and expression which leave familiar faces relatively unaffected (for an overview, see Hancock et al., 2000; R. J. Johnston & Edmonds, 2009).

The other-race effect

The other-race effect (ORE) describes the generally better face recognition skills one has for faces from the own race compared to other races (Chance et al., 1982; Malpass & Kravitz, 1969; Meissner & Brigham, 2001). Although the ORE is widely studied and appears broadly accepted, there are indications that the term is not completely appropriate as there are early indications that it is not ‘race’ that causes the effect, but rather inexperience with certain groups of faces in general, i.e. a racially homogenous face-diet (Mousavi & Oruc, 2020). Thus, the ORE might be a reflection of the low end of the familiarity continuum, where a novel face from another race is the least familiar.

The Face Inversion Effect

The Face Inversion Effect (FIE) is the description for impaired performance on face processing tasks for faces that are shown upside-down (Valentine, 1988). In almost all of the studies

discussed so far, it has been found that inverting the face removes any advantage increased familiarity provides. The FIE provides one of the strongest arguments that there are different processes involved for familiar and unfamiliar faces. For example, in a series of experiments, Megreya and Burton (2006) found they could make a distinction in familiar and unfamiliar face processing based on a subjects' ability to match faces in different conditions of familiarity and orientation. Specifically, they studied the relationship between performance on a face matching task when faces were unfamiliar or familiar, and either presented inverted or upright. They found a correlation between unfamiliar upright and inverted faces, and another correlation between unfamiliar upright and familiar inverted. However, they found no correlation between familiar and unfamiliar face processing when images were presented upright. Additionally, they found no relationship between familiar upright and familiar inverted. These results clearly show that there are different processes used for familiar and unfamiliar faces, which is why it can be difficult to design a task to compare familiar and unfamiliar faces in one task.

Social perception

So far, differences between familiar and unfamiliar faces have been discussed, but it is important to note that there is also processing that can be performed regardless of familiarity. For example, for unfamiliar faces people are still able to recognize expression (e.g. Vicki Bruce, 1986), and categorize age (e.g. Willis & Todorov, 2006) and make trustworthiness judgments (e.g. Palomares & Young, 2018).

However, there have also been differences found in social perception, for example an effect of familiarity in faces where gender was difficult to categorize (Bruce, 1986). There is also a strong effect of familiarity if the task is to match 2 photos and there is a change in one of the social constructs mentioned above, leading to a relatively lower recognition rate of unfamiliar faces compared to familiar faces (e.g. Hill & Bruce, 1996; Patterson & Baddeley, 1977; Roberts & Bruce, 1989). Additionally, some forms of social perception appear to become part of the representation of familiar faces, such as expression. For example, Kauffman & Schweinberger (2004) found that the recognition of familiar faces was faster for moderately positive expressions compared to more intense happy or angry faces. No such effects were found for unfamiliar faces. They argue that the moderately positive expressions are more commonly portrayed (for famous

people in the media), thus it is likely that these more typical expressions become part of the internal representation.

1.5. Facial variation

Section main message: A brief overview of perceivable variation in faces.

One major contribution to the progression of the field of face recognition is the proposal of the “multidimensional space framework” by Valentine (1991) that describes an organizational space in which faces are represented as a location defined by its value on multiple dimensions. I will discuss this model in greater detail in section 1.7 ‘Exploring face representations through (psychological) models’, but will focus this section on the one major aspect this framework did not cover: what visual information is used for the organization of the space. This was made clear by the author himself, stating “These aspects have not been specified simply because there is as yet insufficient evidence available to enable informed choices to be made.” In this section, I discuss the relevant literature that has emerged since Valentine’s proposal (1991) to review what knowledge is available that can guide choices for the dimensions in face space and what information is still missing.

A secondary aim of this section is to provide an introductory overview of perceivable variation in order to compare human and artificial face recognition, in the second part of the thesis. Techniques such as deep learning use all available information in an image and humans - presumably - do not (as this would not be in line with the free-energy principle discussed in section 1.2, Friston, 2009, 2010; Sengupta et al., 2013). Understanding the differences in information used in both systems can help interpret findings and guide model choices. I will discuss the usefulness of deep learning techniques in understanding human face recognition in greater detail in Chapter 7.

To explore potential dimensions of human face space, it must be clear that there are not infinite possibilities, they are limited to all perceivable facial variation. The question “What visual facial information is involved in human face recognition?” has been tackled in quite many ways and in multiple disciplines of science, such as neuroscience, computational cognition and

experimental psychology. As discussed above, facial variation is often studied in an attempt to explore factors that influence face processing, such as familiarity (Hancock et al., 2000), the Other Race Effect (ORE; Byatt & Rhodes, 2004), or the face inversion effect (Mckone & Yovel, 2009; Rossion, 2008). This review is not focussed on these phenomena and their related performance differences but simply aims to get a clear picture of what facial variation is found to be perceivable.

Structural versus pictorial information

In face research, experiments are often designed to ensure that *structural* information is used in the face recognition tasks and not *pictorial* information. Structural information is described as information on the structure of the face that is essential to distinguish it from other faces, and pictorial information is the information that is associated with the depicted specific instance, such as lighting and resolution (Bruce & Young, 1986). This chapter focusses on perceivable manipulation of the input at whatever level, thus, pictorial information was included in this overview. A multitude of studies have indicated that people can perceive face changes caused by illumination (Bruce & Langton, 1994; A. Johnston et al., 1992), hue (Kemp et al., 1996), orientation (Moses et al., 1996), viewing distance/image size and head pose (Hill et al., 1997; O'Toole, Edelman, et al., 1998).

Metrics

A common manipulation to faces is an alteration in spacing between features, often referred to as *configural* information. This type of manipulation is easily done because it only requires the displacement of a face component (a definition of 'face component' is given in the next section). This line of research has shown that people can perceive changes in the distances between the left and right eye, nose and mouth, and eyes and nose (Brooks & Kemp, 2007; Leder & Bruce, 2000; Leder & Carbon, 2006). Additionally, there are studies that categorize types of configural information, for example, in vertical and horizontal displacement, with similar findings (Goffaux & Rossion, 2007). These studies commonly are in the context of the face inversion effect, which is often used to study configural information, because inverting a face impairs the processing of configural information (e.g. Malcolm et al., 2004; as discussed in more detail in Section 1.4. Differences and similarities between familiar and unfamiliar face perception).

Features/local/component

In terms of facial variation, changes to facial components have been a research topic for a long time. In this review, we use the word ‘component’ to describe the parts that make up the face, so the mouth, nose, eyes and eyebrows. In other literature, this is also referred to as a (facial) feature or part. I chose for the word ‘component’ because I believe it is the least ambiguous one. However, in any of the studies discussed and experiments presented in this thesis, it is important to ensure that the component is understood and defined as clearly as possible.

Components can be divided into internal and external: internal is most often defined as nose, mouth, eyes and eyebrows while outside of the outline of the face is considered external. As discussed in section 1.4, there is an increased focus on internal over external features for familiar faces compared to unfamiliar faces.

The literature on manipulation or sensitivity to changes in components is quite large, however, it remains unclear what a change to a component actually entails. To start with a very simple inconsistency in literature, studies on the effect of changes to the eyes could mean changes to a property of the eyes (such as eye colour, Chaby et al., 2003), swapping the eyes for entirely different ones (Abudarham et al., 2019) or swapping the eye region that includes the eye brows (Riesenhuber et al., 2004). Especially the latter two are difficult to interpret in terms of potential dimensions in face space, as completely swapping out a component changes both colour, shape and often also configural information. Additionally, some argue that the eyebrows should be treated as a separate component that has a significant contribution in recognition (Sadr et al., 2003) which would indicate that all the properties of the eyebrow have their own dimensions separate from the eye. This still leaves the question of what should specifically be considered the eye and the eyebrow. This question is the same for all components: what are the spatial borders that define the area? When changing components, there is the obvious risk that configural information is also changed. Thus, the question on what spatial borders define an area is a relevant question when studying the effect of transforming a component. Some studies in the context of the face inversion effect centred the changed feature in the old position, which should supposedly avoid change to the configural information. (e.g. Malcolm et al., 2004). However, this solution assumes the centre of the position matters more than the edges of the component, while the processing of configural information and how we define a feature are too unclear to be confident in this solution.

Another approach in defining components is by defining the dimensions in Valentine's face space (1991), discussed below in section 1.7. For example, Abudarham and Yovel (2016) created a novel method to explore what critical dimensions are used in facial identification. They selected 20 features that were rated by participants in a series of experiments in order to collect feature values and similarity ratings. This study shows that people can perceive variation in: lip thickness, hair colour, eye colour, eye shape, eyebrow thickness, ear protrusion, forehead height, hair length, eye size, skin texture, jaw width, eyebrow shape, nose size, nose shape, skin colour, face proportion, cheek shape, chin shape, eye distance and mouth size. Details of this study will be discussed in section 1.6.

Traits, descriptives and expressions

The previously described research discusses what facial information varies within and between identity. As discussed above, people can also use visual facial information to make social judgements, as we can perceive traits (e.g. trustworthiness, Palomares & Young, 2018), descriptives (e.g. age Willis & Todorov, 2006) and expressions (e.g. Vicki Bruce, 1986), even without familiarity of a face. Social perception of faces has been often treated as separate from identity recognition. However, studies have found influences of familiarity on social perception (e.g. Claypool et al., 2007; Li et al., 2019). Nevertheless, it is difficult to know if the internal visual representation associated with a familiar face is what causes the influence, as the influence could equally well be of a subjective basis due to the semantic knowledge on that person. The relevance of understanding social perception in our attempt to understand face recognition remains unclear, as it remains unclear what visual information is used for the perception of these higher-order constructs. Very limited research in the deconstruction of visual input has been done, but a more recent approach, the 'Bubbles' technique, will be discussed in the next section.

The order of the headings in subsection 1.5 is not random, as all the different types of facial variation discussed are ordered on a scale from low to higher order processing, as far as possible based on the current evidence. The suggested scale is intuitive in the sense that features reflect low level visual information, such as spatial frequency and contrast, on one end of the scale and higher order social perception on the other end of the scale. However, the relationship between the two ends and what constitutes the rest of the scale is far from intuitive.

1.6. Two ends of a scale – measuring individual differences, deficits and enhanced performance

Section main message: This section is meant to give a brief idea of what the performance of people with prosopagnosia and super-recognizers is like and why it is difficult to measure face processing abilities. Most tasks measuring face recognition abilities have a binary response and/or are focussed on differentiating different subprocesses of face processing, making it not directly clear what differences in visual facial representation between familiar and unfamiliar faces are responsible for differences in performance.

Many of the approaches that are commonly used to measure face processing abilities focus on or are developed for one particular population: Individuals with a set of deficits specifically in face recognition, known as prosopagnosia. More recently, there has been more interest in measuring performance of individuals on the other end of the scale, with enhanced face processing abilities, named ‘super-recognizers’. Discussing the insights into face processing that the study of both populations has provided would be far beyond the scope of this chapter. Therefore, we limit the discussion to some brief definition of prosopagnosia and super-recognizers, and what we learned from these two populations specifically about the internal representations of faces. I will end this section by discussing the most commonly used tasks in measuring face processing abilities.

Prosopagnosia

Prosopagnosia is a disorder in which people suffer from an impairment that is specific to face recognition (and does not appear to affect other domains, such as object recognition; Corrow et al., 2016). Obviously, the study on the effect of prosopagnosia specifically on the face representation is even more limited than the research on face representation in the general population, but there are some studies on the topic. Due to the limited literature, studies on both acquired (following brain injury) and developmental (innate) prosopagnosia are considered here.

Some studies on the effect of prosopagnosia on face representation show an intact face space (e.g. Nishimura et al., 2010; Susilo et al., 2010). However, this is also limited to what tests were used, which is discussed further (in context of the general population) in section 1.7 “Measuring face processing abilities”. Other studies, however, have found differences in (the use

of) face representations in people with prosopagnosia. For example, Caldara and colleagues (2005) found in a case-study that the patient's representations of familiar faces seemed to be less robust than those of controls, as she consistently needed more information to reach recognition. Additionally, she used less of the information around the eyes to identify familiar faces, but rather used the lower half of the face (which is considered sub-optimal information). Although it is unclear why, it does suggest there is a relationship between the amount of information used in the internal representation and the 'robustness' of those representations.

Super-recognizers

Compared to people with prosopagnosia, super-recognizers are on the opposite end of the scale, they are people who have enhanced face recognition abilities (Ramon et al., 2019; Russell et al., 2009). The first cases of enhanced face recognition abilities were described by Russell et al. (2009), in which four participants showed performance far above average on the 'before they were famous' test and the Cambridge Face Memory Test. This increased performance was interpreted as evidence for the existence of people with extraordinary face recognition abilities and were subsequently dubbed 'super-recognizers'. However, similar to the people with prosopagnosia, the expression of their face processing abilities can present in many different ways and as a group super-recognizers are quite heterogeneous (Ramon, 2021). Since super-recognizers as a defined group is a relatively recent phenomenon, most research focuses on mapping out their face recognition abilities and exploring how broad this enhanced performance is. However, recent studies have been done that specifically explore the visual information super-recognizers sample, namely very recent work by Dunn et al (preprint; 2021).

Dunn et al. (2021) compared the visual sampling of 10 super-recognizers in an attempt to get a better understanding of the cognitive and perceptual processing that led to their enhanced face processing abilities. To explore visual sampling during a learning task, they used a 'spotlight' that had apertures of different sizes, i.e. only a part of the face was visible through such a 'spotlight' and the amount visible was manipulated. Their findings show that super-recognizers only out-perform 'typical' viewers if they could see at least 36% of the face. Additional analysis was used to explore qualitative difference in gaze behaviour between typical viewers and super-recognizers, which showed that super-recognizers focus less on the eyes, more on the nose and explore more compared to typical viewers. The finding that super-recognizers

focus more on the nose has also been found in a study by Bobak et al. (2016) that compared the eye-movements of DPs and super-recognizers. The findings of Dunn et al. (2021) indicate the encoding of faces for super-recognizers is where the enhanced ability arises, and the authors suggest that super-recognizers should be better considered to be super-learners.

The research on super-recognizers has also led to more emphasis on individual differences in the general population. So far, I have discussed performance differences in general, but not how these are measured. What follows is a brief overview of the different paradigms and outcome measures that are used to measure face processing abilities in order to give a sense of the types of data related to human face recognition research.

Tasks: Face matching

A common task to measure face processing abilities is of course the basic face matching task in which at its core is the comparison between two images. The two images compared could be two different original images, or one original image compared to an altered image. As mentioned before, many of the tests to measure performance were initially targeted at people with prosopagnosia, but later used to explore face processing in a more general population. I have thus far spoken about ‘face recognition abilities’ and ‘face processing’ as the tests often reflect several different subprocesses of face recognition at once, such as (Ramon et al., 2019; Ramon & Gobbini, 2018). The tests most commonly used to measure face processing abilities are the Cambridge Face Memory Test (Duchaine & Nakayama, 2006) and the Glasgow Face Matching Test (Mike Burton et al., 2010). The outcome measures in face matching tasks are percentage correct, reaction time, misses, false alarms and related measures such as d-prime, and area under the curve. For a more complete overview of the mentioned tasks and others, please see (Herzmann et al., 2008; Noyes & O’Toole, 2017; Ramon, 2021; Ramon et al., 2019).

Tasks: Bubbles technique

Very few attempts have been made to directly quantify and measure the visual information we use in face recognition. Perhaps the most popular one, is the ‘Bubbles’ technique. This technique covers the image of a face with a mask and partly reveals information through Gaussian Windows dubbed 'Bubbles', located randomly across the face. The aim of the technique is to quantify the amount of visual information that is being used by comparing task performance to the sampled information/number of bubbles (Gosselin & Schyns, 2001). The information that was linked to a correct response can be considered as the ‘effective’ stimuli’. This technique was then used to derive the effective stimuli in three different categorization tasks, in which participants were asked to categorize identity, gender, or expressiveness. In a follow-up study (Schyns et al., 2002), the authors looked at how many bubbles were used to correctly identify 75% of the trials in each category. They found that the most information was necessary to categorize identity, in which participants needed an average of 33 bubbles, while they only needed 20 for gender and 15 for expression.

The Bubbles technique becomes more informative when the input image is manipulated, such as the manipulation of the spatial frequency in the original paper. Manipulations to the visual information present in the face are an increasingly more commonly used technique within face recognition research, due to technological advances that allow for easier image manipulations. One study that explored differences in external representation by manipulated visual information was performed by Abudarham and Yovel (2016), whose work has been fundamental for this project. Thus, I will discuss this study in greater detail to provide the background for Chapter 2 - Beyond binary face recognition task: The effects of familiarity on sensitivity to subtle face changes.

Tasks: Reverse engineering

Although the experiment in Abudarham and Yovel (2016) can still be considered a matching task in essence, it provides a useful framework to explore the internal representations of faces using altered images in a face matching task. This work is an application of Valentine’s face space (1991), which I will discuss in greater detail in the next section, “Exploring face representations through (psychological) models”. In order to explore the dimensions that make up face space, the

authors (2016) explore what features are ‘critical’ for identification, i.e. the features that, when changed, alter the perceived identity.

Abudarham and Yovel (2016) first formulated a concretely defined face space that consisted of 20 dimensions, which will be referred to as features from here on. Subsequently, the faces were assigned values by participants for each of these features: participants were shown 100 faces they could scroll through (so multiple at the same time) and were asked to rate particular features from -5 to 5. They were asked to rate based on all faces, and could also change their previous scores if they wished to do so. Each participant rated 100 faces on 4 features. Based on the z-score of these ratings, a face vector was created. These vectors were then used to calculate the distance between faces by calculating the L1 norm, which is the sum of the absolute differences between the two vectors. To validate the chosen features, two groups of 10 pairs were created, one group each for the largest and smallest face space distances. Another group of participants (who had not participated in the original rating task) was tasked to rate pairs of faces on similarity and the results showed a correlation between perceptual similarity judgments and their face space distances. Subsequent experiment used this face space to alter identity and specify the most relevant features.

The approach of Abudarham and Yovel (2016) is based on procedures in machine learning, where the input of a classifier is systematically changed in order to find what information is actually used in the process, hence the term ‘reverse engineering’. As the systematic change of the input is the most crucial part of this technique, I will now discuss how the transformation used by Abudarham and Yovel (2016) to ‘move’ faces in face space in more detail. The most extreme values per face vector tell you what feature is most distinctive per face. These features were selected and then replaced by a feature that was furthest away from it in face space, i.e. the maximum change possible within the mapped face space. A total of 30 faces were changed, with a varying number of features changed. The number of features depended on whether or not it still looked natural/ not odd-looking, which was rated by yet another new group of participants.

These transformed images were then used to create pairs in a matching task that had three conditions, Same, Changed and Different, in which a pair was made up of a base image and alternative image of the same person, the alternative image with a changed base image of the same person, or two base images of different people respectively. Participants were presented

with these pairs and asked whether or not these pictures belonged to the same person, with a response scale ranging from 1 - definitely the same person and 6 – definitely different people. The size of the change (perceptual distance) was rated by a new group of participants in a similar procedure as describe in the second paragraph of this section. Abudarham and Yovel (2016) found that the size of the changes they made in terms of perceptual distance were correlated with the perceptual change, further validating the face space Abudarham and Yovel (2016) used.

As should be evident from the discussion about the Abudarham and Yovel (2016) study so far, is that the rating of features to measure perceptual distance is an important part of this framework. In addition to the described experiments, the authors explored whether the inter-rater agreement could be used as a measure for perceptual sensitivity (PS): if a feature was rated more consistently, it must mean it is easier to perceive than features with less consistent rating, indicating the feature is more important for discrimination. The features that turned out to have high PS were indeed the features that showed a strong correlation between perceptual similarity and face space distances, where the low PS features showed a very weak and insignificant relationship. Thus, high PS features appear to have higher discriminative power than low PS features.

Although the original 20 dimensions have been chosen relatively arbitrarily with some apparent guidance of literature, the authors state that the correlations between perceptual similarity and their face space distances can be taken as a validation of the 20 chosen dimensions. However, this does also mean that there are dimensions not in their current face space that could also be relevant. Although the terms ‘dimension’ and ‘feature’ were used interchangeably in the paper by Abudarham and Yovel (2016), it remains unclear if these features were single dimensions or a collection of dimensions. Their method of producing a change was to swap whole face components, such as eyes or hair. Some ‘dimensions’ in their face space could therefore represent visual information that is coded separately. For example, for the dimension/feature hair colour, the method actually resulted in a change beyond just hair colour, and also changed properties such as hair length and forehead height.

The transformation to the face by ‘adopting’ features of other faces fits the purpose, as the framework by Abudarham and Yovel (2016) is clearly stated to explore the visual information that is related to face identification and altering the face in that way will define what features are relevant. However, it provides little insight into how large a change to the original

features would need to be to move an identity out of its face subspace. Although not mentioned explicitly, it appears the authors assumed no changes to any of their features would be large enough to shift identity, as they object to measuring perceptual sensitivity by making changes to the face by altering original features, stating “...this type of measurement is not useful for studying face identification because face identification is about comparing features across different faces.” In other words, they argue that transforming a feature in a face by altering original facial features is not useful as a transformation if the goal of the study is to explore face identification. However, this remains to be seen, as the dimensions of face space are unclear, thus it is unclear if the range of a dimension might be large enough to change identity. This is the theme that is explored in the psychological experiments throughout this thesis.

In sum, while there is a lot of data available from face perception tasks, the transformations used and what it says about the internal representation we hold of faces are often limited. For example, a face is either correctly identified or matched in a two-alternative forced choice task, or incorrectly identified. A binary task can tell us what happens below and above a threshold of recognition, but not what happens at the boundary of where an identity is no longer recognized. It is exactly this space that can tell us more about the internal representations we have of faces, as it provides an insight into what is important identifying information and whether each face is equally influenced by change. It seems unlikely that we know every face equally well and that every unfamiliar face is equally unfamiliar (think of unfamiliar people that look similar to somebody you know), but it remains unclear how this translates to differences in internal representations. Different levels of a rich representation built over time would fit with the capacity of visual working memory (VWM), which is limited but can be allocated dynamically (Ma, Husain, & Bays, 2014). Thus, to gain a greater insight in face perception (and to fully make use of the benefits that deep learning approaches can provide), face recognition tasks with a more complex, but controlled, transformations are necessary. Exploring such stimuli will be the main aim of the psychological side of my PhD project.

1.7. Exploring face representations through (psychological) models

Section main message: Two major influences in face perception research are the formulation of the Bruce and Young (1986) model and the Multidimensional face space framework (Valentine, 1991) as they led to unifying existing research and to many new testable hypotheses and useful implementations.

In the following section, I will discuss two psychological theories of facial representation and two implementations of these theories that have been of major influence for the progression of face perception research, and greatly influence how we think about face processing currently.

The Bruce and Young (1986) model

Although face recognition was an active research field, until the 80's few unifying theories were proposed to account for the perceptual and cognitive processes in face processing. Bruce and Young (1986) set out to connect the few theories that were proposed and extend them to include more aspects of face processing. This resulted in the famous 1986 model, which was the first to provide such a complete integration and overview of the face processing knowledge accumulated so far. However, the true strength of this model is not its descriptive nature of face processing, it is the predictions that follow this framework and the way it has led to new insights in face recognition. I will now describe the model and some other influential work it inspired.

Bruce and Young (1986) aimed to describe face recognition in a very broad sense, in order to capture all related perceptual and cognitive processes associated with face processing, stating "We use the term recognition here in a broad sense, covering the derivation of any type of stored information from faces." The authors formulated 7 types of information that are derived from a face, namely, pictorial, structural, visually derived semantic, identity-specific semantic, name, expression, and facial speech. 'Pictorial' describes information on image properties, such as lighting and photo-resolution. 'Structural' describes the facial information we use to distinguish them from other individuals, and this is the type of information the authors propose we use for familiar face recognition. 'Visually derived semantic' information refers to social perception, such as age and gender categorization which can be done solely based on the image. 'Identity-specific semantic' information refers to information you hold on the person you cannot derive from an image, such as occupation or where you know them from. 'Name' is just the

person's name but is coded separately from identity-specific semantic information, because of existing disorders where there is a specific inability/difficulty in name retrieval. 'Expression' information relates to visually derived judgement on what emotion someone is showing. Lastly, 'facial speech' information refers to the facial information that affects our interpretation of heard speech, such as in the McGurk effect (McGurk & Macdonald, 1976). In this thesis, I focus on pictorial and structural information. For the sake of brevity, I will only explain the parts of the Bruce and Young model (1986) that are related to familiar face processing of identity information alone, leaving out the description of the model for expression analysis and facial speech analysis.

The first stage Bruce and Young (1986) describe is that of structural encoding, which consists of a view-centred description and a more abstract description of the visual facial information, such as configurational information and feature-specific information, dubbed expression-independent information. This latter type of description is fed into the so-called 'face recognition units' (FRUs), which is a space in which descriptions of all the faces known to a person are stored in units. In this space, the expression-independent information is compared to the FRUs and the strength of the signal is related to the similarity of that comparison. The strength of activation can be influenced by non-visual identity-specific information. For example, if you see someone that resembles a colleague, the information on an identity of where you know each other from, such as the workplace, would increase the strength of the signal. A similar relationship is with a component described in a general sense as the 'cognitive system', which represents general non-identity associative memory that can increase the strength of the signal. In the given example, this could mean the workplace is the university and the associative memory could be information about the building you work in. This cognitive system also interacts with what Bruce and Young (1986) named 'directed visual processing', which allows for task-related information to be formed and used, such as holding the distinctive features of a friend's face in mind when you are searching for them in a crowd.

The path of activation in the model described in the previous paragraph is used for familiar faces. For unfamiliar faces the model starts with the same view-centred and expression-independent description of the face but it cannot go to FRU, as no information is stored for an unfamiliar face. Thus, it goes directly to the directed visual processing components, which in collaboration with the cognitive system can be used to complete tasks such as matching

unfamiliar faces. Thus, according to the Bruce & Young (1986) model, the difference between familiar and unfamiliar lies in the lack of a representational unit and all processes associated with that FRU that could boost the activation signal.

The Interactive Activation and Competition Model

Burton, Bruce and Johnston (1990) further explored the Bruce and Young (1986) model with the use of an interactive activation model. Although the paper explored effects of semantic and identity priming in addition to distinctiveness, I will only discuss the distinctiveness findings as this effect stems from differences in visual information in the FRUs and that fits within the scope of this review.

Interactive Activation and Competition (IAC) models were developed in the early 80's and are essentially clusters of units, with bidirectional excitatory connections between units in different clusters that each have their own weight (Burton et al., 1990). The authors created a cluster for FRUs, one for highly identity-specific semantic information and one for associative semantic information. Faces were represented by a hypothetical pattern of activation across a set of input units. This activation functioned as input to the other units, thus spreading across the rest of the network, although all other individual units could also be activated directly by the experimenter (for full specifics on the model, please see Burton et al., 1990; McClelland & Rumelhart, 1980), as the following discussion will solely focus on the distinctiveness manipulation). Within each cluster, units are interconnected with inhibitory links. To test the effects of distinctiveness 6 additional clusters with 6 units were linked to the FRUs cluster, each cluster representing a facial feature (in this experiment they used hair, eyes and noses as an example, but note that it is just hypothetical). When a face was distinctive, it was connected to face-units that other faces were not connected to, while typical faces would have face-units in common with other faces. Thus, distinctiveness was defined by the number of overlapping components, in which a low number of overlapping components is distinctive. They found that more distinctive faces reached the threshold of activation of the highly identity-specific semantic information earlier than typical faces. In other words, more distinctive faces are easier to identify. The reason for this is the lack of competition: the interconnection between units in a cluster is inhibitory, thus competition causes a lower net activation of a unit, which results in a longer time to reach the threshold.

Multidimensional face space (Valentine, 1991)

In both the Burton and Young (1986) model and the IAC model (Burton et al., 1990), the authors made clear that there was little attempt made to specify the details of a facial description used in such a system. In an effort to unify findings on distinctiveness, inversion and race, Valentine (1991) proposed a multidimensional organizational space in which faces are represented, dubbed face space.

Briefly explained, the idea behind face space (Valentine, 1991) is that research on distinctiveness, inversion and race all indicate the influence of knowledge of the population of faces a person encountered so far. Thus, he proposed a framework that can encode a face as a point in a multidimensional space that is based on previously encountered faces. The dimensions in this space represent features that are used to encode a face. The space was proposed to be Euclidean in the example, indicating that the distance between points in this space relate to similarity of faces: the more similar a face, the closer it will be in face space. However, this was more to illustrate the principle as the author did mention that the “Euclidean metric will almost certainly be an over-simplification”.

Although some suggestions are given on what these dimensions could be, no attempt is made to specify these in the model (Valentine, 1991), though many properties of these dimensions are proposed. Each dimension is a feature that has a certain variation within the population and this model assumes that variation is represented by a normal distribution. Face space has an origin, which reflects the central tendency of all the different dimensions. Since similarity is represented by distance between points, the density around the origin of face space is highest and decreases as distance from the origin increases. Additionally, the normal distribution of all features mean that typical faces are close to the origin, and distinctiveness increases as the distance to the origin increases. The origin is based on experience with faces in a person’s lifetime, suggesting the origin changes over time and is adapted to faces encountered.

In this paradigm (Valentine, 1991), face recognition happens in a two-step process: a face must be represented in face space as an n-dimensional vector, and then this must be compared to other representations that are stored in face space. Error and noise should be thought of as an area surrounding the coordinate of the point representing the face. Although Valentine does not explicitly make the connection, his proposed face space could be considered as the “structural encoding” within the Bruce and Young (1986), as the face space described can represent both

familiar and unfamiliar faces. Additionally, no predictions are made about what the differences between a familiar and unfamiliar face in face space would be.

From Pixels to People

Although the IAC model provides insight into how the different types of information in the Bruce and Young (1986) model could influence each other, the authors do note that “Work on these phenomena must focus more precisely on the nature of the information passed into the face recognition system.” The face space model also does not include more details on what visual information is needed in order to form a face vector or what processing is needed to achieve that (Valentine, 1991). To explore what type of perceptual representation could serve as an input for face recognition, Burton, Bruce and Hancock (1999) expanded the previous models with a perceptual front-end.

Burton, Bruce and Hancock (1999) used an adaptation of the IAC model and added image processing before the model to gain a better understanding in what image processing is needed to make an activation that can activate FRUs. For this purpose, they used principal components analysis (PCA) to create image vectors of faces. PCA was first applied to face recognition in the early 90’s by Kirby and Sirovich (1990), and by Turk and Pentland (1991).

PCA is an attractive candidate to use, because it allows for face images to be represented in very few dimensions while still retaining enough information for comparisons. Additionally, PCA is a relatively straightforward and simple technique to represent images of faces. PCA is very sensitive to location differences in images, such as size (different total surface) and position. However, Craw (1995) found that when faces are morphed to a standard shape, such as the average, these ‘shape-free’ faces can then sensibly be subjected to PCA. Subsequently, Hancock et al. (1996) expanded on this approach by using the deviation from the average in landmark facial point locations to derive a shape representation. This made it possible to explore the contribution of shape and texture, in which texture was defined as the “color information and the fine-scale features unaffected by the shape averaging”.

Models of face recognition were largely divided between perceptual and cognitive approaches. As mentioned above, Burton, Bruce and Hancock (1999) expanded the previous models with a perceptual front-end to explore what type of perceptual representation could serve as an input for face recognition. I.e. they created a model that had both a perceptual and

cognitive component. They (Burton et al., 1999) found that PCA serving as a perceptual front-end combined with the IAC model can model various multi-modal effects, such as distinctiveness, semantic priming and repetition priming. One point raised explicitly by the authors (1999) about the use of PCA and the successful modelling was the important implication that image-based representations of faces are enough to account for certain effects in face recognition, instead of the need for a more abstract type representation that some researchers previously assumed. The PCA representation of this model can deal with variation in expression but it will not deal well with varying view-points. However, even if the specific procedure of PCA might not reflect human-like mechanisms, the concept of an image-based representation of a face has proven to be a very insightful finding.

Face-space-R - Lewis 2004

Another implementation that has led to insight in the nature of face space has been provided by Lewis (2004), who expanded on previous versions and interpretations of Valentine's (1991) face space by proposing a revised version that he dubbed 'face-space-revised', 'face-space-R' for short. One of the characteristics of this revised version of the model is that the ease of recognition is explicitly defined. Ease of recognition is calculated based on all the exemplars that inhabit face space and was dependent on three factors: error of encoding and similarity to a target's first and second nearest neighbour. Although face-space-R does not provide more insight in the dimensions of face space, it does provide great insight on how the presence of exemplars can change face recognition performance.

The model itself had three parameters that could be explored: the total number of faces the face space holds (as known individuals), the number of dimensions of the space, and the size of the error. In face-space-R, a face representation is activated based on the similarity between the probe and a particular face representation (of a known face). The recognition of a face is based on the activation of an exemplar, but also that of others: the response to a known identity is the activity of an exemplar minus the activation of all other exemplars. The model was able to account for the effects of (apparent) other race faces, caricature, distinctiveness and familiarity. An especially important addition in this model is the encoding error: an extra component in the formula that allows for the strength of activation to be increased or decreased. Lewis (2004)

shows how effects of presentation time can be mimicked using different values of error: the longer the presentation duration, the lower the encoding error.

Lastly, learning and familiarity were represented by a strength factor, that could increase or decrease the activation of an exemplar. The strength factor would have a value between 0 and 1, where 0 represented a face never encountered before, and 1 meant highly familiar. The strength vector is a cost-efficient solution: each identity/exemplar's strength is dependent on the local density, i.e. how many other exemplars surround it and how closely. Thus, faces in a more densely packed part of face space, will need a larger strength vector to be recognized and this reflects the difficulty in learning typical faces (typicality paradox, Burton & Vokey, 1998). Lewis (2004) also states that there could be a scenario where even the highest strength value is not enough for recognition, and then more dimensions will need to be added in face space. How a face gets a larger strength factor, i.e. is learned, or how encoding error and strength factor relate is not addressed in this paper.

Although great insights have been gained from the conceptualizations of face recognition, all of the models discussed above have stated explicitly that they left out the nature of the dimensions that make up face space. The models explain some effects of familiarity, but it remains unclear to what extent there are differences in the actual representation of familiar and unfamiliar faces. As I have discussed in this review, there is an abundance of research that shows differences in face processing between familiar and unfamiliar faces, but also some research that indicates we use the same information for certain face identification tasks for familiar and unfamiliar faces. In this thesis, the main aim is to gain more insight in what visual information we hold in the representations of familiar faces and how these representations compare to the representations of unfamiliar faces.

1.8. Structure and context of this thesis

Hopefully it became evident in the review that there is a wealth of research on face recognition and a difference between familiar and unfamiliar faces, but it still remains unclear what internal face representations we use. The aim of this thesis is to investigate what internal representations of faces we have and how these are different between familiar and unfamiliar faces.

Structure of the thesis

For the sake of clarity, my thesis is divided in two parts: part I will cover the relevant psychological literature and behavioural experiments and in part II I will cover current technological advances in the field of deep learning and the computational experiments.

In Chapter 2, I examined a new paradigm in which participants could alter certain features in a gradual scale, to explore accuracy of an internal representation. Experiment 1, I checked if the percept between familiar and unfamiliar faces was the same by showing the original on-screen and asking participants to match the image to the original. In Experiment 2, I repeated Experiment 1, but added a delay between the original and the response image, so that face memory needed to be utilized for the task. In Experiments 1 and 2, the features that were changed were distance between the eyes and the eye brows, width of the mouth, lip thickness and brow thickness.

Chapter 0 presents Experiments 3 and 4, that explore the relationship between perceived similarity and image-based spaces in order to define well-controlled stimuli that could create multi-dimensional changes by deploying PCA. Little research has been done on the relationship between perceived similarity and image-based spaces which makes it difficult to quantify face transformation, leading often to the use of verbalizable transformations (such as changing eye colour, skin colour or apparent sex). By comparing the effect of multi-dimensional changes on perceived similarity and 3 different image-based spaces, these experiments aimed to assess the usefulness of PCA transformations in the exploration of internal representations of the face. Chapter 4 repeats Experiment 4a, but with a smaller size of transformation.

Whereas Chapters 3 and 4 focuses on the relation between familiarity and similarity, Chapter 5 explores the relation between familiarity and sensitivity to change. Two experiments are presented, using a two-alternative forced-choice task and two different magnitudes of change. The different magnitudes of changes were used to explore effects of cross-identity versus within-identity changes.

Chapter 0 explores the boundaries of identities further, by exploring the relationship between familiarity and the consistency in perceived change. In this Experiment (6), different sets of components were changed and presented to participants in a 1-to-4 array. Participants were then

asked which of the four changed pictures they thought had changed the appearance of the person the most.

In the literature review of part II (Chapter 7), I will discuss the work that has been done using deep convolutional neural networks to explore human face recognition. Specifically, I discuss the most relevant contributing factors in the development of the field of computational cognition, and why in particular deep neural nets are suitable to explore human face recognition. I end this chapter by reviewing the handful of studies that applied deep neural nets to explore the internal representation of faces.

Chapter 8 describes the three computational experiments in which I mimic familiarity by training on certain identities and comparing the representational spaces of familiar (trained) and unfamiliar (untrained) faces. A pre-trained FaceNet (Experiment 7a and 7c) and VGGFace (Experiment 7b) will be used to quantify changes to an image of a face. In Experiments 7a – 7c, the Euclidean distance will be used to quantify changes. In Chapter 9, I compare the representational spaces using different distance measures, to find which outcome measure best reflects a human behavioural pattern.

In the general discussion, I will summarize my main findings and identify several future directions that could build on the framework provided within this thesis.

Context of the research

My doctoral thesis bridges psychology and computing and allows for direct comparisons between DNNs and humans. Experiments with human participants and DNNs, while technically challenging, allow a better understanding of the workings of both human face recognition and DNNs. This is because, unlike with human brains, we can control exactly what information DNNs are trained on (i.e., what visual experience they had) and we can explore how they use this information compared to humans. For people, techniques such as neuroimaging can provide only limited insight into the information processing, while the exact pathway of information fed into a DNN can – theoretically – be followed throughout. Comparison with DNNs can thus serve as a proxy for our information processing.

Since this work has implications beyond the topic of familiarity and is relevant for our exploration of representations used by artificial systems, it is important to briefly consider the

ethical implications of the research presented in this thesis. Any research in face recognition can be considered a contribution to the ongoing debate about the ethical and social issues surrounding the use of artificial face recognition (AFR). Even if I do not set out to improve AFR systems/algorithms, it is not unreasonable to think that these findings could be applied in some fashion to improve the evaluation and/or performance of an algorithm. Additionally, the descriptions I have used for some transformations have meaning for some social constructs, for example, ‘race’. Because I do not use these descriptions in any way that is related to the social construct, it is important to clearly state the limitations of the findings, and on occasion, what the findings do not mean. For example, I do not use the term ‘race’, but I only discuss ‘apparent race’, as this work ought not be considered as arguments in a larger social debate about concepts such as ‘race’ and ‘sex’. The focus of the proposed work is the comparison of effects of familiarity between humans and DNNs, therefore, it is not an evaluation of the performance (in terms of accuracy) of any AFR system.

However, highlighting differences between human face recognition and AFR powered by deep neural networks (DNNs) can be considered as exposing flaws and biases of AFR. Knowing these limitations is helpful for raising awareness of the potential risks, which is important for any policy and regulatory decisions made associated with these techniques.

2. Chapter 2 - Beyond binary face recognition task: The effects of familiarity on sensitivity to subtle face changes

Chapter main message

Recently, Abudarham, Shkiller and Yovel (2019) found there are no differences in feature hierarchy between familiar and unfamiliar faces in a face identity task. Although the feature hierarchy is not affected by familiarity, memory accuracy of the features in the hierarchy might be. Theoretically, familiarity could aid or disrupt face perception. Familiarity could aid the comparison of two faces by using the memory of a familiar face to enrich the image currently in visual working memory (VWM) with more details to compare. Alternatively, this 'enrichment' has the potential to disrupt the comparison, by overwriting VWM content of the presented image with the visual information retrieved from the memorized face. The current study proposes a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the accuracy of face perception. The effects of familiarity on sensitivity to two types of adjustments, eye brow and lip thickness (high perceptual sensitivity according to Abudarham et al. 2019) and eye distance and mouth width (low perceptual sensitivity) were studied. Two versions of the experiment were used, the first with an on-screen comparison (face perception) and the second with a delayed comparison (face recognition). Our findings indicate that perceptual sensitivity but not familiarity has an influence on the accuracy of our face perception, but neither influence the accuracy of recall of the selected manipulations.

2.1. Introduction

It is well known and accepted that people process familiar faces differently than unfamiliar faces (e.g. Hancock, Bruce, & Burton, 2000; Johnston & Edmonds, 2009). However, it remains unclear what the specific differences between familiar and unfamiliar face recognition are. Recent studies showed no differences in feature hierarchy between familiar and unfamiliar faces in a face identity task. (Abudarham et al., 2019; Abudarham & Yovel, 2018). In the current study, the quantitative differences in the feature hierarchy that are related to familiarity will be explored.

Most of the studies that explore differences between familiar and unfamiliar faces focus on the visual input, often asking the participant to make a same/different judgement, compare/rate or to spot a difference (e.g. Abudarham, Shkiller, & Yovel, 2019; Abudarham & Yovel, 2018; Clutterbuck & Johnston, 2004). While there have been attempts to understand the sensitivity to face changes by using staircases (Brooks & Kemp, 2007), the majority of studies of the effect of familiarity are still based on binary response options or subjective similarity ratings. Although these studies can still provide us with valuable insight, the limited response options allow us equally limited conclusions about the actual remembered content of a particular face. More recently, a non-binary paradigm was used by (Sandford & Burton, 2014) to explore configural processing in which participants were instructed to adjust image ratios until it looked 'right'. Although focused on a specific type of information (configural), the use of a continuous scale provides more insight in the content of facial memory in a relatively low number of trials. Quite surprisingly, no advantage was found for familiar over unfamiliar faces, although this effect was found in brand logos. To go beyond exploring configural information, a paradigm is needed in which feature information can be presented on a continuous scale. In the current study, I explore a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing measurement of the accuracy of feature information in face perception.

In every type of task in which visual information is encoded and/or compared, either to an internal representation or to an externally displayed source, visual working memory (VWM) is involved. VWM is a component of working memory. VWM has a limited capacity (Luck & Vogel, 1997; Ma et al., 2014), meaning there is only a finite amount of visual information at a certain moment in time that we can encode when perceiving a face. Additionally, there is a

theory that the default of the human brain is to minimize the costs involved in any process (e.g. the Free energy principle, Friston, 2009, 2010; Sengupta, Stemmler, & Friston, 2013), making it unlikely that the visual system will keep collecting visual information on features indefinitely. From VWM, the information moves for further processing. There is a bidirectional transfer of information between working memory and long-term memory (for a review, see Burgess & Hitch, 2005).

Theoretically, familiarity could aid or disrupt face perception. Familiarity could aid the comparison of two faces by using the memory of a familiar face to enrich the image currently in visual working memory (VWM) with more details to compare. Alternatively, this ‘enrichment’ has the potential to disrupt the comparison, by overwriting VWM content of the presented image with the visual information retrieved from the memorized face. The current study proposes a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the precision of face perception. It has been shown that loading information from LTM into VWM can enhance task performance (e.g. semantic priming, for a review see Bruce, Dench, & Burton, 1993). Thus, a more familiar face with a better representation of that face in memory will contain more information that can be loaded into VWM and that should allow for a better comparison to what is presented externally.

Although Abudarham et al. (2019) show that the order of their feature hierarchy is not affected by familiarity, this does not automatically mean that the memory precision of the features in the hierarchy are not. For example, perhaps the eyes are more important than the nose for both familiar and unfamiliar faces, but the details remembered on the eyes could still be greater for familiar versus unfamiliar faces. This is in line with a study by O’Donnell and Bruce (2001), who found that sensitivity to changes made to the eyes was enhanced for faces that were familiarized. The current experimental design allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the precision of face perception. The effects of familiarity on sensitivity to two types of adjustments were studied, eyebrow and lip thickness (high perceptual sensitivity according to Abudarham & Yovel) and eye distance and mouth width (low perceptual sensitivity). If there is any effect of having an internal representation present in LTM, precision could increase if loading the representation aids or decrease if the internal representation ‘overwrites’ VWM content.

In addition to exploring the effects of familiarity and PS on subtle face changes, the effects caused by experimental design were also explored. In much face research it is common to test between subjects, especially when the question involves familiarity. For example, to test effects of familiarity on a same-different task, it would not be uncommon to test in the UK using UK celebrities and perform the same test in another country where the celebrities are considered unknown. Although a between-subject design with the same stimuli excludes image-specific effects, it introduces noise caused by subject-related factors, such as individual differences and cultural effects. However, a within-subject design has the opposite problem, as you avoid the issue of effects by individual characteristics but have a potential effect caused by stimuli since the same stimuli cannot be familiar and unfamiliar. To explore any differences brought on due to within and between subjects testing, all participants were included and divided in two groups: a ‘mixed’ group of people who were familiar and unfamiliar with between 4 and 8 of the 12 shown identities and a ‘single’ group with two subgroups of participants who knew or did not know (nearly) all of the identities shown. The one category group can be used to test between-subjects and the two-category group can be used to test within-subject. Subsequently, the results of the two groups can be compared to explore differences in precision related to experimental design.

In the current study, the effects of familiarity on sensitivity to two types of adjustments will be studied, eyebrow and lip thickness (high perceptual sensitivity according to Abudarham et al., 2019) and eye distance and mouth width (low perceptual sensitivity). To explore the effects of familiarity on sensitivity to face changes, participants were asked to perform a face matching task. Instead of a commonly used 2AFC, an experimental design that allows participants to gradually adjust the appearance of a face to match the original image was used, allowing us to measure the precision of face perception. In order to compare our findings of face perception to several studies Abudarham et al. (2019) ran in this paradigm on face identity, a similar sample size was chosen here. I hypothesized that both perceptual sensitivity and familiarity have an influence on the precision of our face perception.

2.2. Methods

Sample size

Due to the shift to online testing and the novelty of a continuous response scale, with an unknown effect size, it was not possible to set the sample size of the first experiment based on power analysis or simulations. Therefore, sample size is based on a rule of thumb that is commonly used for Linear Mixed Effect Models, which is to have a minimum of 30 data points for 30 individuals per group (Hox, 1998). Familiarity is divided in unfamiliar and familiar, thus there are two groups. Therefore, sample size is 60.

Participants

We tested a total of 67 participants (25 males, mean age = 20.7 years, SD = 3.2, age range = 18 - 30 years) of which 32 participants (15 males, mean age = 19.9 years, SD = 1.8, age range = 18 - 26 years) were familiar with 4 - 8 of the 12 shown identities (mixed group) and 35 participants (10 males, mean age = 21.5 years, SD = 4.0, age range = 18 - 30 years) who knew less than 4 or more than 8 (single group). All participants had normal or corrected-to-normal visual acuity. All participants signed an informed consent form after a short explanation about the procedure. Participants received 1 course credit or a monetary reward of £4,- for participation afterwards. The experiment was approved by the General University Ethics Panel of the University of Stirling.

Apparatus and stimuli

The experiment was programmed using E-prime and the image preparation was done using python and Psychomorph (Tiddeman, 2011). The task was displayed on a HP E243 monitor with the refresh rate set at 60 Hz and a spatial resolution of 1920 by 1080 pixels. A common confound in face familiarity research is the unavoidable difference in image properties between the images used in both categories if you use random unfamiliar faces. Aside from the difference in image properties that could be present, one could argue that famous faces are better remembered due to inherent properties of famous people. For example, famous people tend to have a more attractive face than unfamiliar people. One solution for this is for participants to indicate who they know

and do not know and keep the same stimuli for both conditions of familiarity. In order to collect data in which familiarity is variable between groups, 12 different faces of famous youtubers were used. The identities were chosen based on the current top Youtubers combined with student feedback.

We selected two high PS and two low PS features as described by Abudarham et al. (2019), which are lip-thickness and eyebrow-thickness (high PS), and mouth-size and eye-distance (low PS). PS is a measure that indicates the sensitivity to a change in the face and this measure was based on the combination of the outcome of two experiments concerning size of facial features: ratings of facial features and face matching. The specific position of these features in a feature hierarchy are dependent on the results of the specific tagging experiment followed in Abudarham et al. (2019). However, for the current experiment, only a division between high and low PS features is needed and for all mentioned features there is evidence that it will never fall into the other category. For example, changing eyebrows consistently decreases accuracy in face recognition and even if it will not be in the top 5 of the high PS features in our stimulus set, it is extremely unlikely that it will fall below the low PS features in our stimulus set.

Stimuli adjustments were performed using Psychomorph (Tiddeman, 2011). Before the alterations, Adobe Photoshop © was used to remove light reflections in the eyes, as the white pixels provided cues about the correct answer. Alterations to the Psychomorph-templates were done by adjusting the values of all associated face-points. For the sake of reproducibility, we used an automated way of template creation. The script resulted in 21 images in which either eye distance and mouth width or lip and brow thickness were gradually changed. An example of the resultant images is given in Figure 1 and the example of one ‘step’ from the original is shown in Figure 2. The images shown were the images used in the practice trials of the task.

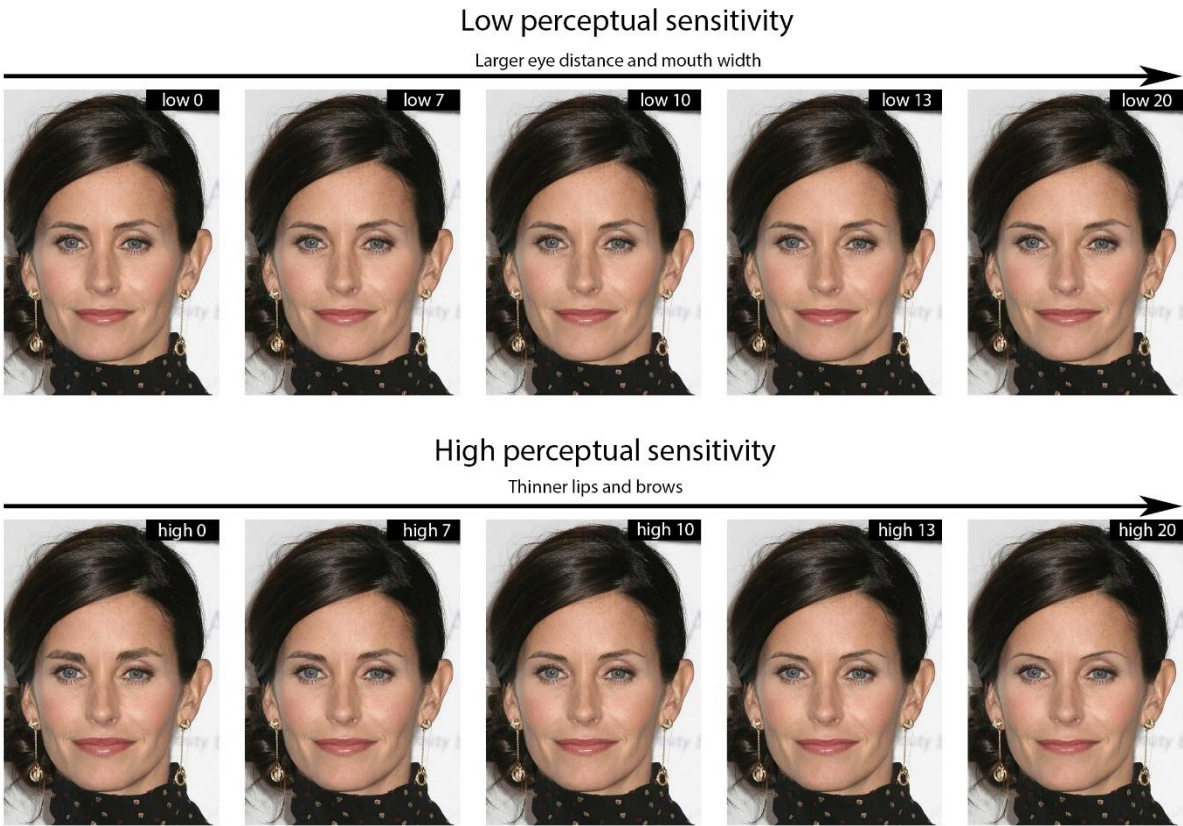


Figure 1 Example of the subtle changes created.

Low perceptual sensitivity



High perceptual sensitivity



Figure 2 Example of 1 'step' from the original.

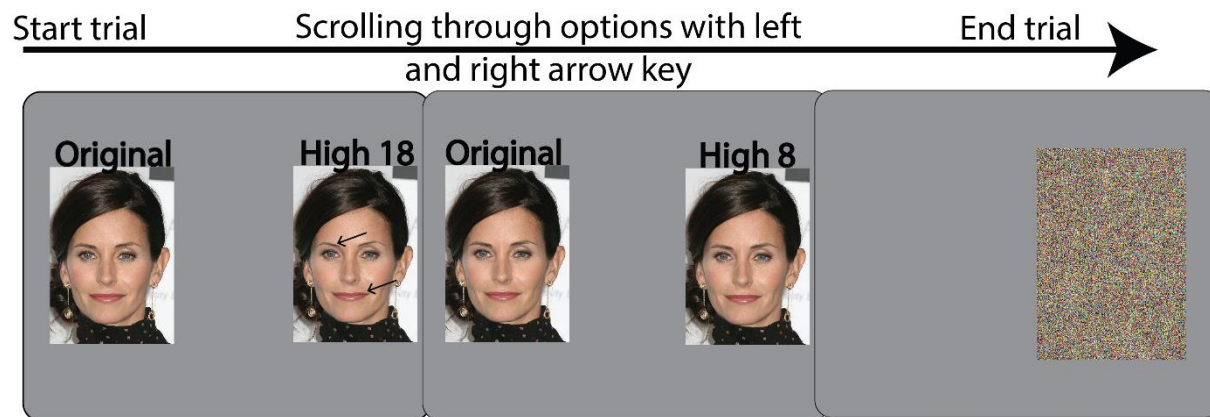


Figure 3 Overview of one trial.

The participant was instructed to make the resemblance as best as they could. The trial would start showing the original on the left and the response image on the right. The response image started at a different value each trial and a variable range of 16 images out of the total 21 images was chosen to avoid strategic scrolling through the images. A mask was presented at the end of each trial to avoid effects of afterimages.

Procedure

The experiment consisted of 2 practice trials and 2 blocks of 48 trials. Trials were blocked based on PS type. Order of PS type and familiarity were randomly selected per participant as a measure of counterbalance. Participants were instructed to answer the question “Try and match these faces as closely as possible. – Make them resemble as best as you can”. The experimenter talked the participant through two practice trials, one for each PS type, to ensure that the participant was able to detect the small difference between each variation of the response image. For each trial, a random range of 16 images was selected out of the total 21 images, to avoid strategic response, such as picking the centre of the range of possible response options. Additionally, the first presented response image would vary in value. For example, in trial one images 1-16 were shown with image 13 as a first suggestion, and in trial two image 4-20 were shown with image 8 as first suggestion. The first face pairs were shown and the participant used the right and left arrow to match both images as closely as possible. Once the participant was satisfied with the result, (s)he pressed spacebar to continue to the next trial. After space bar was pressed, a mask covered the response image for 750ms to avoid feedback. Subsequently, the next trial would

commence, and the participant repeated the same process. And overview of a trial can be found in Figure 3.

Data analysis

We calculated the difference score ('distance' to the original image) for each trial, which is the absolute difference between the selected image and the original image. For example, if the participant selected altered image number 14, the distance to the original would be 3. Participants were expected to show a difference in precision, as represented by a difference in the difference score, between familiar and unfamiliar faces. The effect of PS on precision is well described in previous studies of Abudarham & Yovel (e.g. 2016) and provided a suitable validity check for the functionality of the task. Thus, it was checked if participants indeed performed better in the high PS condition compared to the low PS condition as a way to validate the task. To explore the effects of familiarity and PS, I constructed two linear mixed effects models (a null model and a model containing the variable of interest), for each independent variable and the interaction between both, with the difference score as dependent variable. The linear mixed effect analysis was performed in R (R Core Team, 2019) and *lmerTest* (Kuznetsova et al., 2017). The two models were then compared using an ANOVA. The threshold for significance was set at $\alpha = 0.05$. Normality of the residuals on the models were inspected visually and none of the models showed a violation.

2.3. Results

In the current study, the effect of familiarity and PS on precision in a novel face matching task was investigated. I hypothesized that participants' precision would differ based on familiarity. The descriptive statistics of experiment 1 are shown in Figure 4.

Perceptual sensitivity

The basic null model with as random factor the slope and intercept per participant for the effect of PS was compared to the PS model which was the basic null model with the addition of PS as fixed effect. Our results show that PS affected precision ($\chi^2(1) = 81.01, p < 0.01$), increasing it by distance-steps 0.97 ± 0.08 (standard errors). As hypothesized, high PS features are more accurately perceived than low PS features.

Familiarity

We constructed two models which were compared to explore the effect of familiarity on face memory precision. Our results show that Familiarity does not affect the capability of matching faces ($\chi^2(1) = 0.0042, p = 0.95$). This finding is not in line with our hypothesis.

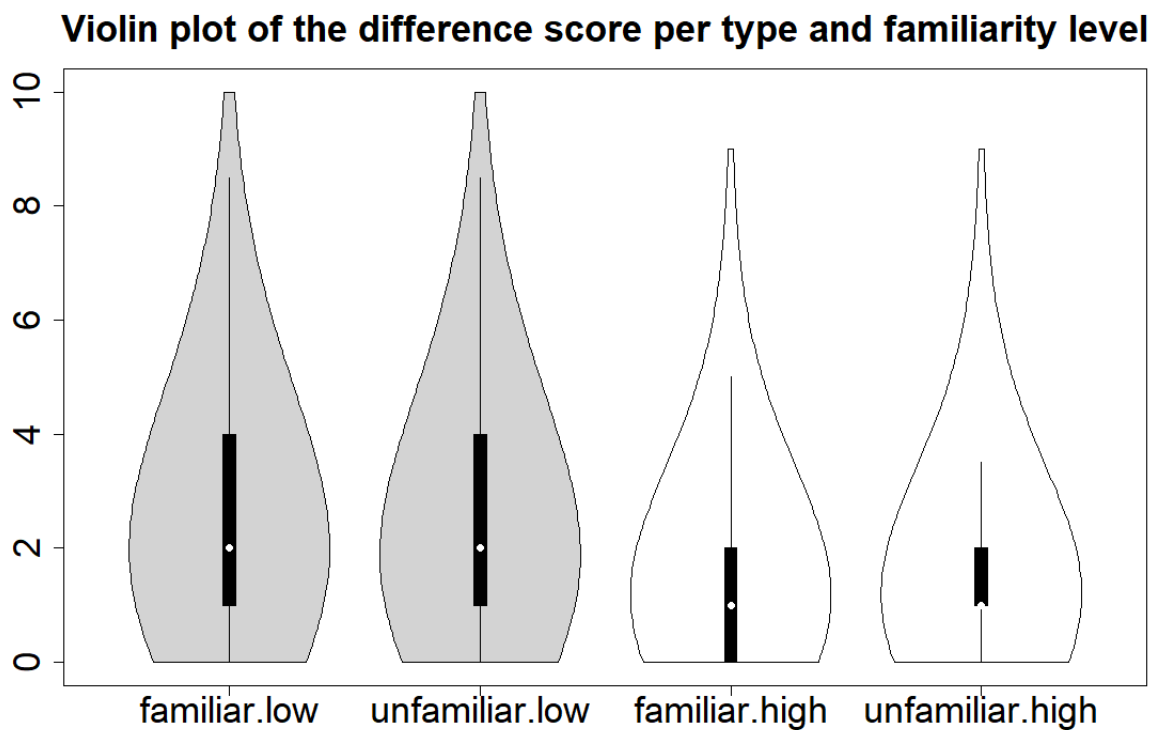


Figure 4 Violin plots of difference score per type and familiarity level.

The white dot indicates the median.

PS and Familiarity interaction

We constructed two models which were compared to explore a possible interaction effect between Familiarity and PS on face memory precision. The results show no interactions between Familiarity and PS ($\chi^2(1) = 2.10, p = 0.15$).

2.4. Discussion Experiment 1

Our results indicate that perceptual sensitivity has an influence on the precision of our face perception, but familiarity does not. These findings provide further support to the recent findings of Abudarham et al. (2019) in which there is no difference found in features used to identify familiar and unfamiliar faces. Although the current findings are clear, these conclusions are limited to on-screen face matching which is similar to the testing conditions of Abudarham et al. (2019). There are a multitude of studies showing the effects of familiarity in face recognition in a variety of experimental designs (e.g. Ritchie et al., 2010). By using low level metric manipulations, the current study indicates that level of familiarity does not alter the perceptual ability of perceiving changes in size. Thus, a possible explanation for the absence of significant results in the current study could be that the effects of familiarity simply occur at higher order than perception or that the nature of the perceptual difference is not related to size estimates (low level perceptual ability).

Alternatively, it could be possible that the presence of both images on-screen promoted a pictorial comparison. The lack of difference due to familiarity could possibly be an effect of the difficulty of the stimuli, as faces contains an abundance of visual information. It is known from concepts such as ‘active vision’ (Hayhoe et al., 1998) that visual information retrieved from eye fixation can be used to avoid the need for a complex internal representation from the external world (the face stimuli in this case). Recently, it has been shown that the cost of memory usage and the cost of the eye movements influence the use of both (Melnik et al., 2018; Somai et al., 2019). The difficulty of comparing two faces on such small differences could have shifted the trade-off between VWM and eye movements towards eye movements, because internalizing the subtle differences would not be cost-efficient. The pictorial strategy would also explain the low medians (1 for high PS and 2 for low PS), as these could be interpreted as the perceptual

threshold in this context. The idea that size is not relevant for familiarity fits with recent findings that the internal representation of a person is based on our prior experience and is idiosyncratic (Ritchie et al., 2018). A possible future direction is to explore other manipulations in the same design. Additionally, it is worth measuring reaction time as familiar faces are matched faster in a binary matching task (Clutterbuck & Johnston, 2002, 2004).

The significant difference found between low and high PS was expected but needs to be interpreted with caution. In the current experiment, only two low PS features were explored, and both were horizontal changes, whilst both high PS features are vertical changes. However, one could argue that the vertical changes are the reason that they are high PS, as it is found that vertical changes disrupt face processing more than horizontal changes (Crookes & Hayward, 2012; Goffaux & Dakin, 2010; Goffaux & Rossion, 2007). Additionally, similar to familiarity, the change was only focused on the size of the feature and PS can be altered in multiple ways. Size was chosen because it can be changed in a rigid and controlled way. However, future exploration is needed to properly define and quantify PS and this paradigm provides a steppingstone for the design of the necessary tasks.

In sum, our current findings support those of Abudarham et al. (2019) that the same critical features are used for identifying familiar and unfamiliar faces. However, these findings only relate to face perception, as the stimuli to compare were on-screen simultaneously. In experiment 2, a delay between the images is added to see if there are effects of familiarity on subtle face changes in face recognition once face memory is involved.

2.5. Experiment 2

The results of Experiment 1 show no differences between the perception of familiar and unfamiliar faces. As mentioned above, Abudarham, Shkiller & Yovel (2019) found that the same features are used for familiar and unfamiliar face identification. The results of Experiment 1 found no evidence to contradict this when there is no memory component involved. However, there is clear evidence of differences in face processing based on familiarity (e.g. Hancock, Bruce, & Burton, 2000) and it is likely that these differences are quantitative in nature (R. J. Johnston & Edmonds, 2009). Since the perception of faces does not seem affected by familiarity,

the next step is to explore differences in internal face representation by introducing a memory component to the task. Since only familiar faces have an existing representation, I hypothesize that participants will perform the task better when shown a familiar face than an unfamiliar face. The increased performance will be evident by a lower difference score for familiar faces than for unfamiliar faces. Additionally, based on the literature (e.g. Johnston & Edmonds, 2009; Ramon & Gobbini, 2018), a more efficient type of processing of familiar faces could be assumed, so a lower reaction time for familiar faces compared to unfamiliar faces is expected.

Pilots

Several pilots were undertaken before the finalization of Experiment

2. The familiarity check was exactly the same:

before the main task, participants shown all the identities and were asked to answer yes if they knew their identity and would recognize them

walking by on the street. The first pilot was almost identical to Experiment 1, but with the addition of a delay between the original and the altered image. To clarify, the current experiment shows the original, followed by a mask, followed by the response image that can be altered. It also measured the reaction time (for an overview, see Figure 5). The original image was shown for 4000 ms as was the mask.

The results of the pilot (N = 10) are shown in Figure 6 and show a performance on the task similar to Experiment 1. This is a very surprising result, as it seems to imply that people remember subtle changes very accurately. However, it could also be explained by a confound in design, if the potential role of spatial working memory (SWM) is considered. SWM is the part of

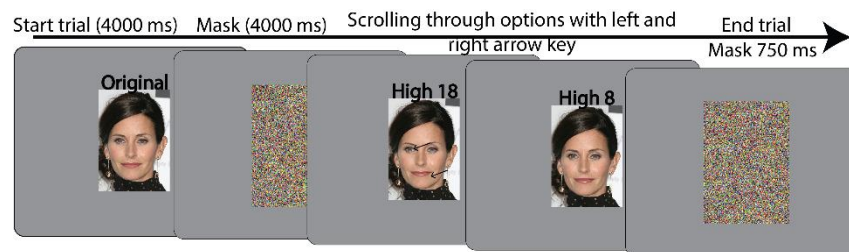


Figure 5 Overview of one trial.

The participant was instructed to make the resemblance as best as they could. The trial would start showing the original for 4000 ms, followed by a mask of the same length and the response image afterwards. The response image started at a different value each trial and a variable range of 16 images out of the total 21 images was chosen to avoid strategic scrolling through the images. A mask was presented at the end of each trial to avoid effects of afterimages.

working memory that is responsible for maintaining information on relevant locations in the visual world. In the current design, the spatial location between the original image and altered image was the same, so this high performance could be more dependent on spatial working memory instead of memory of the face. Additionally, the hard edges could have provided essential spatial information that could be efficiently used in the task. Alternatively, the long exposure duration may have allowed for an effective use of an image-matching strategy (matching to an internal representation of that specific image).

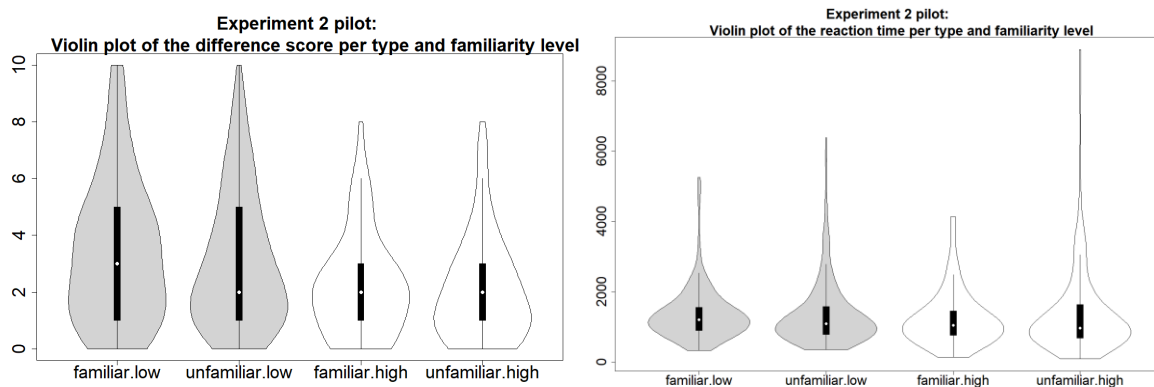


Figure 6 Violin plots of difference score (left) and the reaction time (right) per type and familiarity level in the pilot.

The white dot indicates the median.

The results of and feedback on the first pilot, led to the second version of the experiment, which is shown in Figure 8. In this version, the edges of all the images are blurred and the original image is shown on the left side of the screen and the altered image in the centre of the screen. Two versions with different exposure duration (1000 ms and 2000 ms) were run to explore the effect of a shorter exposure duration. Data of nine participants was collected in each version. The results of the second pilot for both the 1000 ms and the 2000 ms version are shown below (Figure 7).

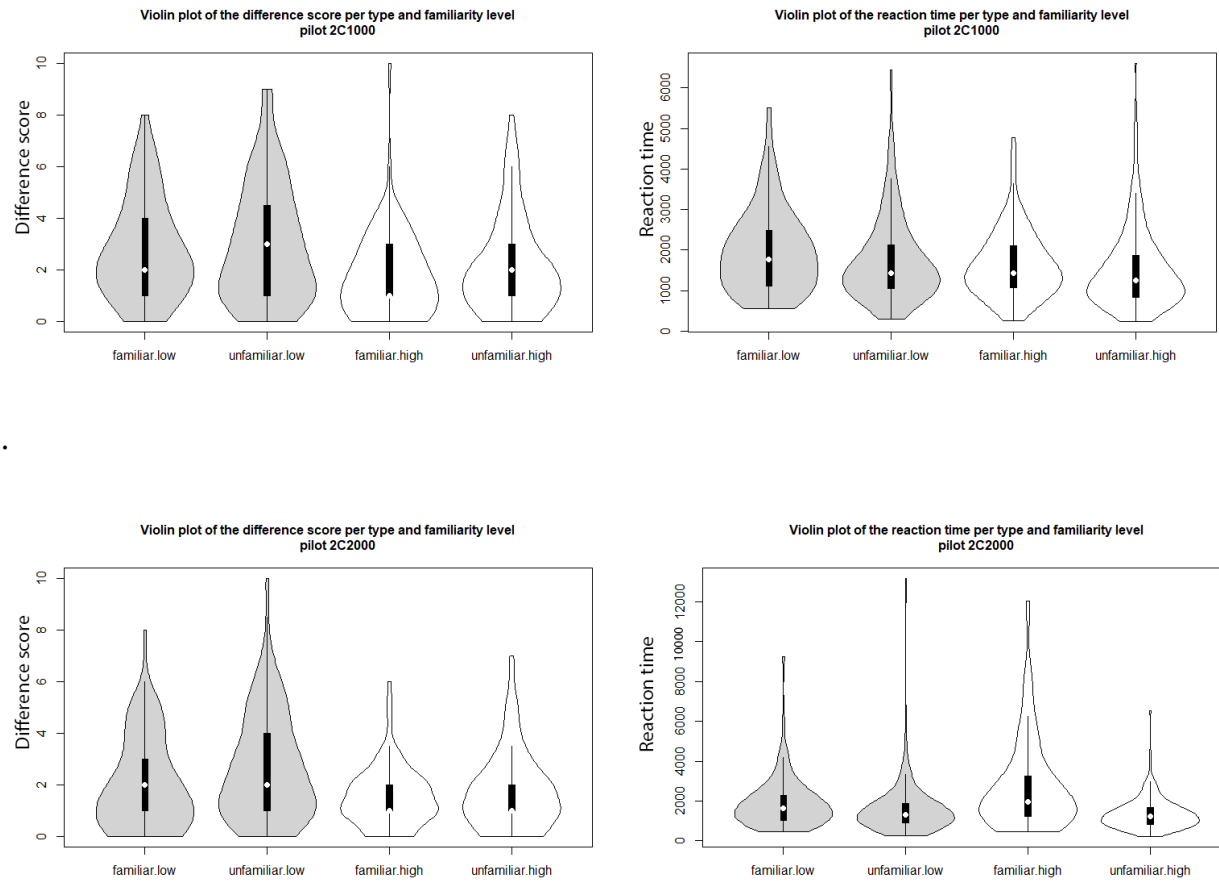


Figure 7 Violin plots

Violin plots of difference score (left) and the reaction time (right) per type and familiarity level for the pilots with 1000 ms (top) and 2000 ms (bottom) exposure duration of the original image. The white dot indicates the median.

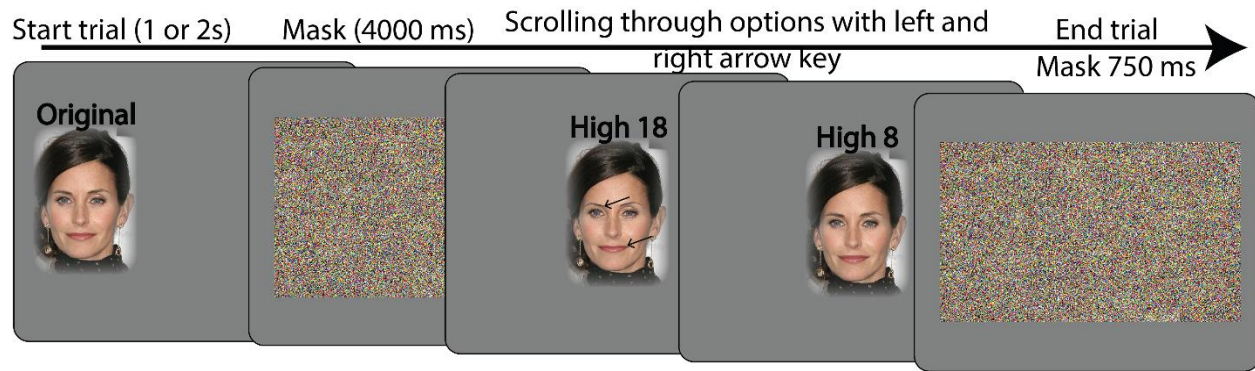


Figure 8 Overview of a trial in the second pilot

in this version the exposure duration of the original image was set to 1000 or 2000 ms and the rest of the experiment remained the same.

As it can be seen, the accuracy remains relatively high in the sense that it is similar to the pilot results of the 4000 ms version. To avoid an image matching strategy and to promote the use of an internal representation of that individual (instead of the shown image), the original image before the trials was removed in the next version of the pilot (N=10). Since there is no original to match to, the task was changes from ‘try and make the resemblance as great as you can’ to ‘make it look right (for you)’. The image is still shown in the beginning as part of the familiarity check, but is not repeated after this instance (for an overview, see Figure 9).

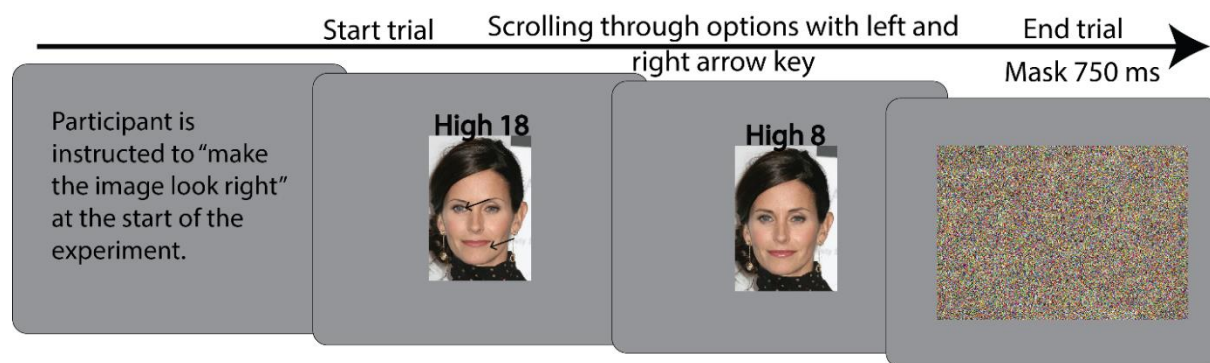


Figure 9 Overview of a trial in the final pilot and subsequent experiment

there is no original image showed (participants see a different image of the identity in the familiarity check) and are instructed to “make the image look right (for you).” The response image remained the same as in the previous experiment.

Although performance remained high (if focused on the size of the second and third quantile and the median), the data appeared to have a higher variability compared to version in which the original photo was shown. However, this is just based on visual inspection, since the sample was too small to run a proper analysis on the data. The results are shown in Figure 10.

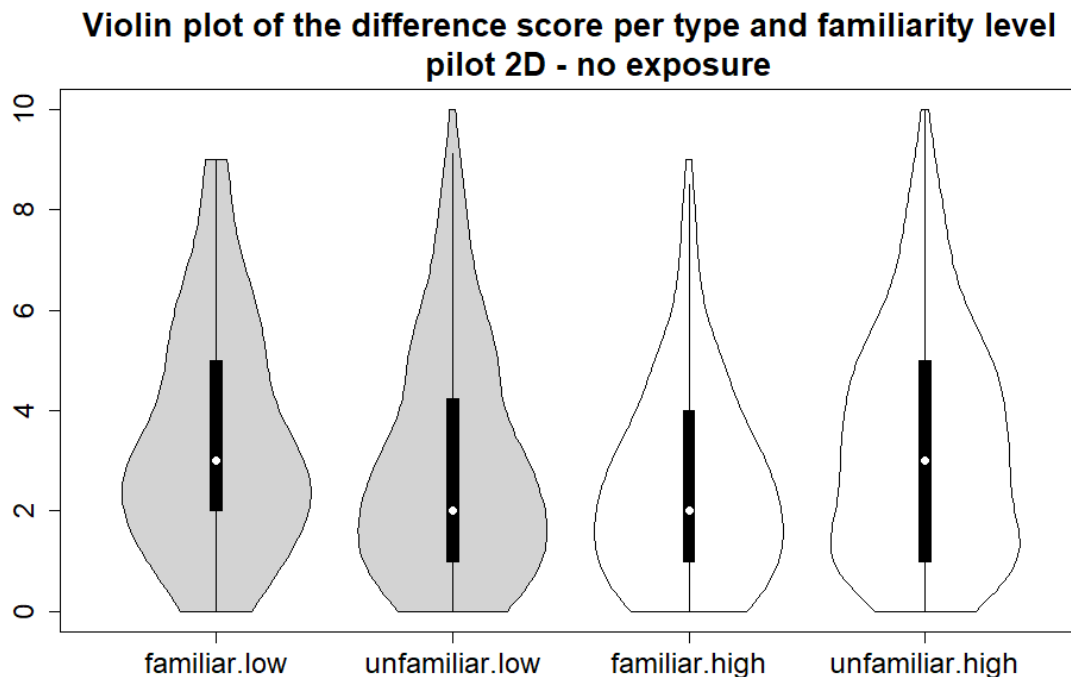


Figure 10 Violin plots

Violin plots of difference score (left) and the reaction time (right) per type and familiarity level in the final pilot. The white dot indicates the median.

The results are easier to interpret when split on level of familiarity, PS and identity. Especially in the high PS condition, shown in the graphs below (Figure 11), you can see that the variability appears to be marginally greater for unfamiliar than familiar faces, as indicated by the length of the quantiles. These results suggest that the current level of difficulty and exposure duration are suitable for exploring familiarity and thus I decided to use this version as a second experiment.

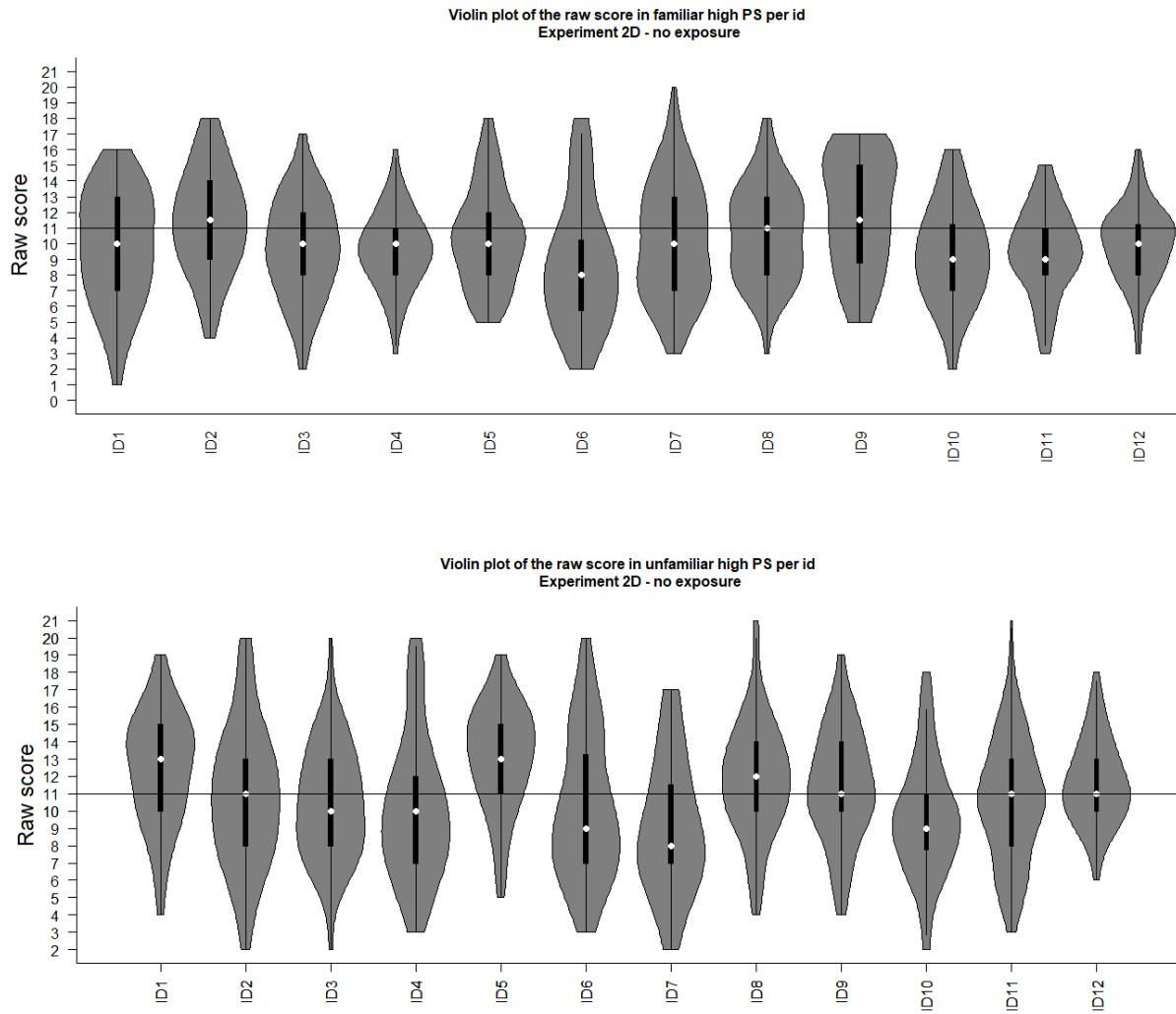


Figure 11 The raw scores in the high PS condition of the final pilot split in familiar (top) and unfamiliar (bottom) per identity.

2.6. Methods experiment 2

Sample size

Due to the novelty of a continuous response scale, with an unknown effect size, it was not possible to set the sample size of the experiment based on power analysis or simulations. The results of the first experiment were expected to give an estimation of effect size, which could subsequently be used in the power analysis for further experiments. However, no significant differences were found.

Therefore, sample size was based on literature for this experiment. Specifically, sample size is set to match that of the binary equivalent of a similar face matching task by Abudarham and Yovel (2019).

Participants

A total of 40 participants took part in the experiment (7 males, mean age = 21.3 years, SD = 5.4, age range = 18 - 43 years).

Apparatus, stimuli and procedure

The experiment is very similar to Experiment 1 but has the following changes in order to explore recognition instead of perception: blurring of the edges of the presented images and there is no original image presented, apart from in the familiarity check. Experiment 2 was performed on the same set up as experiment 1 and used the same stimuli.

Data-analysis

A similar statistical model as formulated in Experiment 1 was used in this experiment. However, instead of using a null model and adding variables in, the ‘maximal type’ of model was used in which all variables are included from the start. The outcome measure of reaction time was included, because differences in reaction time could indicate a difference in processing due to familiarity (e.g. Besson et al., 2017; Ramon & Gobbini, 2018).

2.7. Results experiment 2

In the current experiment, the effect of familiarity and PS on precision of the internal representation was explored. Based on the outcome and related literature of Experiment 1 (R. J. Johnston & Edmonds, 2009; Ramon & Gobbini, 2018), I hypothesized that participants' precision would differ based on familiarity. The descriptive statistics of experiment 2 are shown in Figure 12.

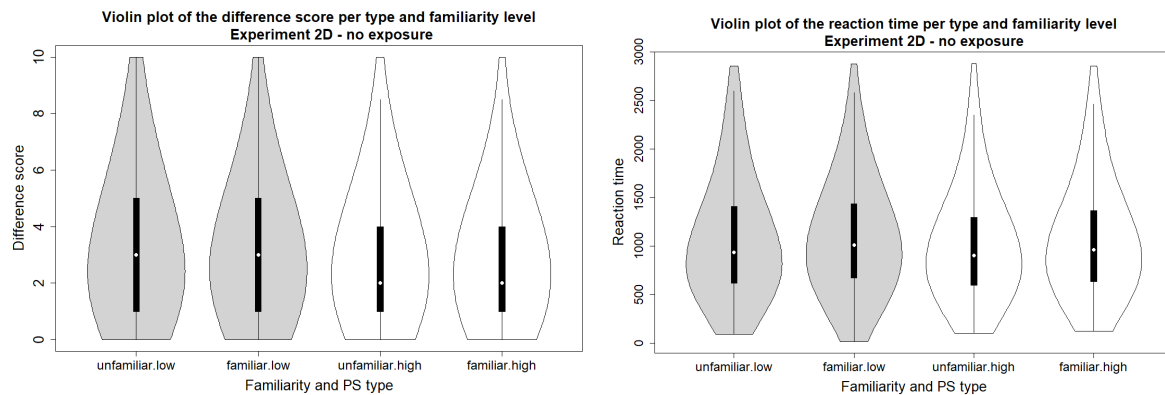


Figure 12 Violin plots of difference score (left) and the reaction time (right) per type and Familiarity level in experiment 2. The white dot indicates the median.

The ‘maximal type’ model was used to analyse the results of the experiment on any effects of Familiarity, PS and the interaction between both. For the difference score, our results show that participants did not perform better on familiar faces than on unfamiliar faces ($\beta = -0.13$, $SE = 0.14$, $t = -0.93$, $p = 0.359$), PS showed a trend but does not significantly affect the capability of matching faces (high compared to low PS: $\beta = -0.35$, $SE = 0.17$, $t = -2.02$, $p = 0.056$) and there was no interaction between the two ($\beta = -0.13$, $SE = 0.22$, $t = -0.58$, $p = 0.563$). For the reaction time, our results show that Familiarity had no effect on performance ($\beta = 14.09$, $SE = 35.88$, $t = 0.39$, $p = 0.697$), PS also does not significantly affect the capability of matching faces ($\beta = -58.61$, $SE = 48.71$, $t = -1.20$, $p = 0.236$) and there was no interaction between the two ($\beta = -1.64$, $SE = 46.73$, $t = -0.04$, $p = 0.972$).

2.8. General discussion

In this chapter, the effects of familiarity on sensitivity to subtle face changes were explored by using a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the precision of face perception. Two versions of the experiment were used, the first with an on-screen comparison and the second with a delayed comparison. Effects of familiarity on sensitivity to two types of adjustments were studied, eyebrow and lip thickness (high perceptual sensitivity according to Abudarham et al., 2019) and eye distance and mouth width (low perceptual sensitivity). The results of both experiments indicate that familiarity does not have an influence on the precision of our face perception. These findings provide further support to the recent findings of Abudarham et al. (2019) in which there is no difference found in features used to identify familiar and unfamiliar faces.

The results of the second experiment are very surprising and not in line with the literature on familiarity (Hancock et al., 2000; R. J. Johnston & Edmonds, 2009; Ramon & Gobbini, 2018). The trend of a main effect of PS in this experiment might indicate that the task has potentially become too difficult if we consider Experiment 1 as a baseline (in which the main effect of PS was very strong with a $p < .01$). If this was the case, the lack of difference and the overall performance would simply reflect noise and guessing. However, since the paradigm is novel, the lack of data to compare the performance to leaves us with an unclear picture of what ‘random’ performance would look like.

Before this outcome can lead to a conclusion about face recognition performance and the absence of any effect of familiarity, it is necessary to check whether this result is specific to the features that were manipulated, i.e. the changes to the eye and mouth region. The current paradigm used very basic and low-dimensional changes to the faces. However, based on literature on the variability within and between the representations of faces (Burton et al., 2016; Lewis, 2004; Valentine, 1991, 2001), it is likely that there are more complex changes relevant for face processing. In animal research on neuronal spiking in face processing, it has been shown that different neurons respond to different types of face manipulations (Freiwald et al., 2017). The metric changes made in the experiment might be processed by lower-level neurons which are not affected yet by familiarity. In the following experiment, high-dimensional (but still

subtle) changes to the face were chosen as a transformation, based on a principal component analysis (PCA).

In conclusion, our study provides a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the precision of face perception. Our current findings support the finding of Abudarham et al. (2019) that the same critical features are used between identifying familiar and unfamiliar faces.

However, in the interpretation of these findings it is vital to understand that the alterations to the faces only reflect a fraction of the variability possible in faces. To gain greater insight in the results of experiment 2 and explore the generalizability of the findings, the same experiment with multidimensionally altered faces was run (experiment 4).

3. Chapter 3 - Exploring perceptual similarity and its relation to image-based spaces: an effect of familiarity.

The work presented in this chapter has been published by Taylor & Francis in Visual Cognition on 17/6/2022, available at <http://www.tandfonline.com//doi/full/10.1080/13506285.2022.2089416>.

Chapter main message

The lack of controlled stimuli transformations is an obstacle to the study of the effects of familiarity on face recognition. Researchers are often limited to verbalizable transformations in the creation of a dataset. An alternative approach to verbalization for interpretability is finding image-based measures that allow us to quantify image transformations. In this study, we explore whether PCA could be used to create controlled transformations to a face by testing the effect of these transformations on human perceptual similarity and on computational differences in Gabor, Pixel and DNN spaces. We found that perceptual similarity and the three image-based spaces are linearly related, almost perfectly in the case of the DNN, with a correlation of 0.94. This provides a controlled way to alter the appearance of a face. In Experiment 4a, the effect of familiarity on the perception of multidimensional transformations was explored. Our findings show that there is a significant relationship between the number of components transformed and both the perceptual similarity and the same three image-based spaces used in Experiment 3. Furthermore, we found that familiar faces are rated more similar overall than unfamiliar faces. That is, a change to a familiar face is perceived as making less difference than the exact same change to an unfamiliar face. The ability to quantify, and thus control, these transformations is a powerful tool in exploring the factors that mediate a change in perceived identity.

3.1. Introduction

In the previous chapter, we explored the effects of familiarity on sensitivity to subtle face changes by using a novel experimental design that allows participants to gradually adjust the appearance of a face to match the original image, allowing us to measure the precision of face perception. Two versions of the experiment were used, the first with an on-screen comparison and the second with a delayed comparison. We studied effects of familiarity on sensitivity to two types of adjustments, eyebrow and lip thickness (high perceptual sensitivity according to Abudarham et al., 2019) and eye distance and mouth width (low perceptual sensitivity). The results of both experiments indicate that familiarity does not have an influence on the precision of our face perception.

Before this outcome can lead to a conclusion about face recognition performance and the absence of any effect of familiarity, it is necessary to check whether this result is specific to the features that were manipulated, i.e. the changes to the eye and mouth region. The current paradigm used very basic and low-dimensional changes to the faces. However, exploring other changes than the ones tested in the previous chapter is challenging, as the lack of controlled stimuli transformations – i.e. the creation of facial variation – is a known obstacle to the study of the effects of familiarity on face recognition. The aim of this chapter is to explore a method for generating variations in the appearance of a face image, and methods for assessing the likely effect on human perceptions of such variations. We then use these methods to test the effect of familiarity on the perception of suitably controlled alterations to faces.

Most research that explores facial variation, uses the approach of reverse engineering: manipulating the visual input to assess whether this content is used in face identity recognition. For example, changing facial components such as the nose, mouth and eyes (Abudarham & Yovel, 2016) or displacing them (Goffaux & Rossion, 2007). This approach is useful but requires well-understood transformations of the visual input to be meaningful. For example, a change in eyebrow thickness is easily verbalized and can be varied systematically, but a change based on 'morphing' - gradually changing from one identity to another - is difficult to interpret in terms of dimensional change. Verbalizable transformations are insightful but are only a part of the complete story.

In order to successfully identify a face under a wide range of viewing conditions, our internal representations of a face need to capture variability in appearance, in order to “tell

people together”, but also tell people apart (Burton et al., 2011; Jenkins et al., 2011). Burton et al. (2016) explored the way faces vary using PCA, and found that the variation is idiosyncratic. One identity may vary in ways that another does not. As an example, someone with a prominent brow ridge will show more variation in lighting on their eyes. Although the exact nature of between and within face variability has yet to be determined, it is clear that this variation goes beyond easy to verbalize features. An alternative approach to verbalization for interpretability is finding image-based measures that allow us to quantify image transformations. Such an approach would need image-based measures that preferably correlate with perceptual space, and well-controlled transformations.

Principal Component Analysis (PCA) was first proposed as a means of face recognition by Kirby and Sirovich (1990), and Turk and Pentland (1991). PCA produces a compact representation of a face but is very sensitive to image properties such as lighting. Nevertheless, using controlled images, PCA will reveal systematic differences between faces, such as gender (O’Toole, Deffenbacher, et al., 1998). The analysis is greatly improved by first morphing all faces to the same average shape and performing PCA on the shape vectors and the ‘shape free’ faces separately (Craw & Cameron, 1991). This separate analysis yields significant correlations with human perceptions of memorability and distinctiveness of faces (Hancock et al., 1996). It can also be used to synthesise novel face images (Hancock, 2000) and by extension, provides a controlled way to produce changes in the shape or surface appearance of face images. It is this ability to produce controlled transformations that is exploited here.

One aim of this chapter is to test the utility of PCA-induced changes in face images to probe human face perception, and subsequently use these transformations to explore the effects of familiarity on face recognition. PCA can produce graded changes and it is to be expected that a large PCA change will produce a bigger perceptual difference than a smaller one. However, it would be helpful to have a separate computational measure of the likely perceptual effect of varying different components or combinations of them. If a purely image-based measure (without a biological premise) correlates with perceptual ratings, it can be used in the creation of stimuli to make a prediction of how a human observer will perceive the image. For example, one could predict whether a certain image transformation is strong enough for the altered face and the original face to be perceived as different from one another. Here, we test three different computational measures of image similarity.

The simplest measure of image similarity is pixel-wise Euclidean distance. For this to have any chance of working for face images, they need to be very carefully controlled. A simple misalignment between two copies of the same image will produce a large difference, while changes in lighting will also result in a large difference that may be barely noticed by a human observer. Because of this, the measure is little tested in face perception. However, since our images are generated by a PCA model that is based on averages and we select the later components which result in filtering out variation such as lighting and angle (for details, see Methods: apparatus and stimuli), they are inherently well-controlled. Meytlis (2011) used carefully controlled images and found an almost perfect correlation between distinctiveness ratings and Pixel Euclidean distance.

A more biologically plausible measure commonly used in computer vision (Lades et al., 1993; Wiskott et al., 1997) and neurocomputational models of face recognition (Dailey & Cottrell, 1999; O'Reilly & Munakata, 2000) is Gabor filtering. Gabor filters are found to be a good representation of how cells in the primary cortex respond to a certain scale and orientation (De Valois & De Valois, 1990; Jones & Palmer, 1987; Ringach, 2002). As Gabor filtering reflects very early processing and not any higher order processing, it would seem unlikely that Gabor distances would relate to perceptual similarity of faces. However, this is exactly what Yue et al. (2012) found: Gabor-based image measures strongly correlate with human performance in a discrimination task with complex figures. For this reason, we included Gabor space to quantify image transformations.

Current artificial face recognition deploys deep neural networks (DNNs) and their face matching ability easily surpasses human performance in many situations. Hancock, Somai and Mileva (2020) compared six state-of-the-art DNNs with human performance on four face matching tasks designed to be challenging for human participants. The humans averaged 73.3%, while the best DNN achieved an accuracy of 98.6%. Despite the difference in accuracy, there were significant correlations between the six DNN similarity scores for a pair of faces and the human accuracy on that pair. This was especially true for matching pairs, where the two images do show the same person. Humans and the DNNs seem to agree on what makes faces look similar. However, the correlation for mismatch face pairs was lower, indicating that humans and the DNNs show less agreement on what makes faces look different. Indeed, the DNNs worked in distinctly non-human like ways, often reporting that faces which differed in apparent race or sex

were the same identity. Given these discrepancies, it is valuable to include DNN similarity measures in the present analysis of controlled image changes.

Hancock, Bruce and Burton (1998) looked at correlations between human similarity ratings and PCA on existing face images. Here, we invert that process and ask humans to rate the similarity of face images that have been altered in a controlled way by varying principal components. We compare the human data with similarity scores obtained from Gabor, Pixel and DNN analysis of the same images. In Experiment 3, we test the effects of single component transformations of both colour and shape, using unfamiliar faces, with the aim of providing better methods for choosing stimuli. In Experiment 4a, we test the effect of multiple component transformations on both familiar and unfamiliar faces. This allows us to test the effect of familiarity on perceptual similarity.

3.2. Experiment 3 - exploring the effects of single component transformations

Background

One major issue in exploring face identity recognition is finding an image transformation that reflects a likely dimension in face space which can also be interpreted in some meaningful way. PCA transformations can be difficult to interpret, as they reflect variation in pixel values that commonly occur together. For example, a PCA component that changes eye colour will also change hair colour, because people with darker eyes often have darker hair than people with lighter eye colours. Thus, verbalizing the change of PCA components can be difficult. On some occasions, a component can be classified after visual review, such as a component that varies apparent masculinity. However, for most components the variation is difficult to classify. In this experiment, we will explore the effect of single component transformations on Gabor, Pixel and DNN space, and (human) perceptual space in order to assess their usefulness in quantifying image transformations.

3.3. Methods

Sample size

Due to the idiosyncratic nature of the image transformations, it was not possible to set the sample size of the experiment based on power analysis or simulations. Therefore, sample size was based on literature and previous studies (with different stimuli). A common group size for such studies is around 30, as averages tend to stabilize around this number. For example, Hehman et al. (2018) explored when averages stabilize (i.e. variation in resampling data points is below a threshold value), and found that most traits stabilize with a sample size of 30 or less.

Participants

We tested 25 participants (11 female, 13 male, 1 other, mean age = 39.7 years, 2 missing values, SD = 16.3 age range = 12 - 74 years. Participants were naive to the purpose of the experiment and took part on a completely voluntary basis without any monetary compensation. The experiment was approved by the local General University Ethics Panel of the University of Stirling.

Apparatus and stimuli

The experiment was programmed through the online platform Qualtrics and participation was allowed via PC, laptop, phones and tablets. Face image stimuli were generated from a PCA-based face model, written in Matlab (MATLAB, 2019). The PCA space was calculated from 72 eye-aligned images, each an average of 10 images of a famous (mostly Caucasian) female. Psychomorph (Tiddeman et al., 2001) was used to create the averages. Use of average images helps to reduce variations due to lighting and viewpoint. Even so, the first few components encode image properties such as lighting and head movements, so we selected components 10 – 16 to alter. Three different types of transformations were used: shape-only, colour-only and both. To explore the effect of the size of the transformation, two transformation sizes were used: 3 and 6 SD. The SD for each component was scaled according to the variance in the (72- eye-aligned) original images. This gave 42 pairs in total: 7 shape, 7 colour and 7 both, two transformation sizes each. For each pair a different face image was generated by setting the first 20 shape and colour components to a random normal value, mean zero, SD 1, though truncated at +/- 2SD to prevent faces looking too strange. An example of the transformation is shown in Figure 13. In addition to the PCA transformations, three single feature transformations derived from Independent Component Analysis (ICA) were added. However, these were purely exploratory; a

discussion is given in supplementary materials. The images produced were all 420 by 595 pixels, size in the experiment was browser and screen dependent.

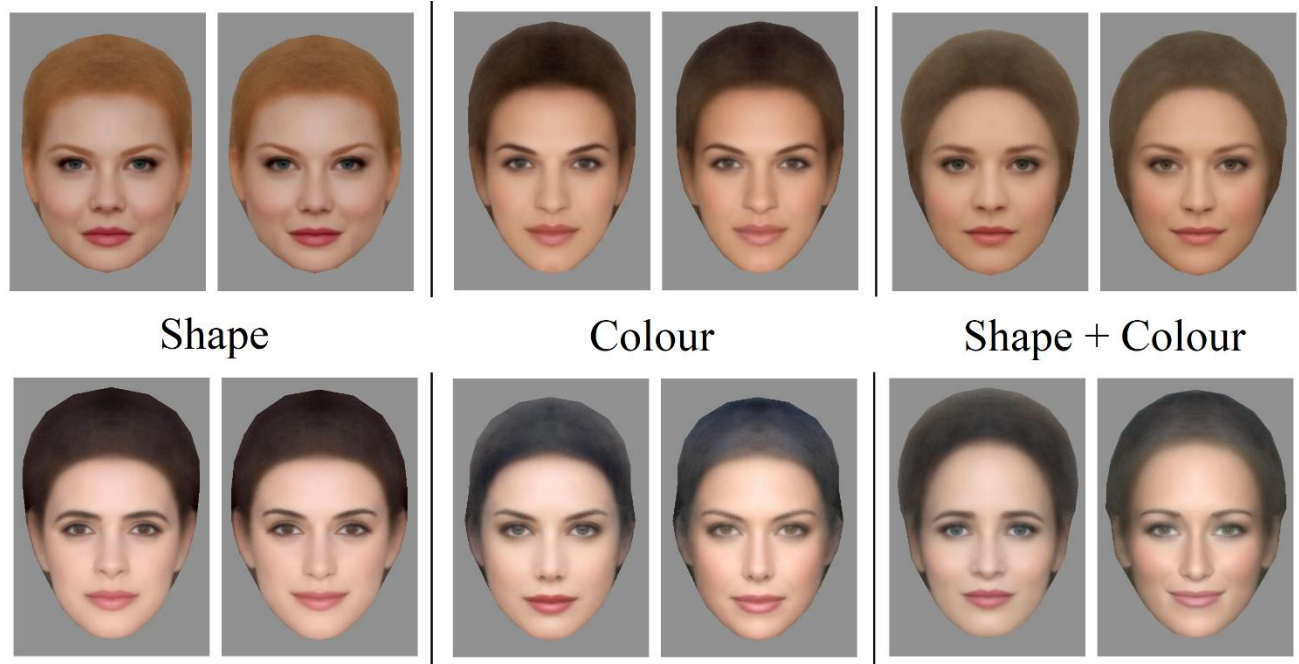


Figure 13 Example of the transformations used.

On the top row, 3SD transformations and on the bottom row 6SD transformations. For each size of change, the left column of each pair shows one end of the scale which is -1.5 and -3 SD from the original respectively and the right column is the other end of the scale which is +1.5 and +3 SD. The left column shows shape transformations only, the middle column shows colour transformations only and the last column shows shape and colour transformations. The base image for each pair was generated at random, as described in the text.

Procedure

Participants were shown an information sheet and then gave consent online. The face matching task consisted of 45 trials (14 alterations for three conditions and the 3 exploratory ICA transformations) that were presented in a random order. In each trial, two faces were presented along with the question “Could these be two pictures of the same person? 1 – certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. An example of a trial is shown in Figure 14.

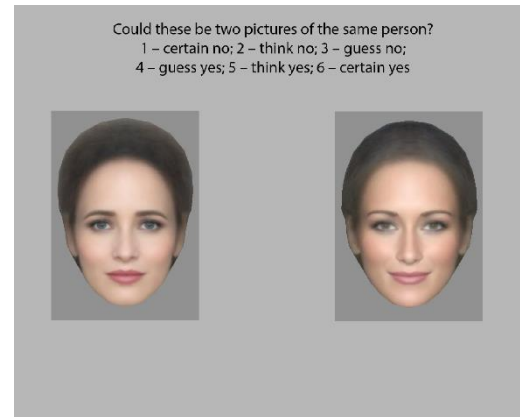


Figure 14. Example of a trial

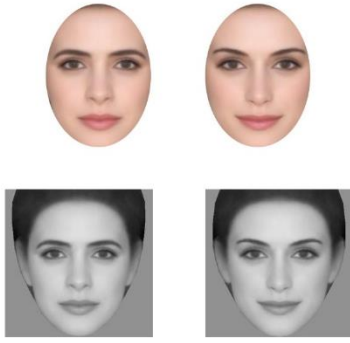
The three computational spaces

Gabor

We used the Gabor Features in Signal and Image Processing Toolbox (Kämäräinen et al., 2002b, 2002a), written in Matlab (2019). This requires a 256 square input image, so the original images were cropped to just below the chin, resized and converted to monochrome, Figure 15. We used the default settings of the toolbox, which computes phase and magnitude for 8 orientations and 5 scales across a 10x10 grid, giving a vector of 8000 values for each image. The difference between two images is given by the Euclidean distance between the two vectors, calculated using the norm function in Matlab (MATLAB, 2019).

Pixel

A simple way to compare two images is Euclidean distance in pixel space. Since our images are all aligned by the eyes, pixel distance might plausibly capture the perceptual changes caused by variations purely in the colour space. However, shape transformations will result in different face features being compared, or face pixels in one image compared with background in another. Meytlis (2011) normalized each image to the same average grey level and cropped them to a tight oval. Since we are comparing pairs of images that differ by adding a principal component



that is inherently zero mean we did not normalize the colour values, but did crop tightly around the face with an oval 378 x 277 pixels, see Figure 15. The difference between images was computed by concatenating the RGB vectors and calculating the Euclidean distance between them using the norm function in Matlab (MATLAB, 2019).

Figure 15 Example of cropped images, differing in shape component 10. Top row, for Pixel comparison, bottom row for Gabor.

DNN

The DNN used was an experimental system from the University of Surrey. It produces an output vector of length 512. The similarity of the two images was given by the cosine of the angle between the two output vectors. Identical images give a similarity of 1, with smaller values being less similar. This is in contrast to the Euclidean distance of Gabor and Pixel spaces where larger numbers mean less similar.

3.4. Results Experiment 3

We treat our participant responses, about whether the two faces look like different people, as a similarity scale, with 6 being the most similar. On a 6-point scale, 3 was “guess different” and 4 was “guess same”. Thus, the boundary for same or different identity was at 3.5. We report the agreement between the average human classification and that offered by our 3 computational spaces. The deep neural network has a specified decision boundary at a similarity score of 0.4. For the two image-based approaches, we computed the minimum distance between all 72 of the images used to generate the PCA space. This provides a conservative estimate of ‘different’: any pair less similar than this threshold should be different people but it is possible that images of different identities might be more similar than this threshold. It should be obvious that a threshold could be set that more accurately describes the data based on the findings. However, this is exactly why we wanted to find a rule of thumb for setting a threshold: creating a suitable

stimulus set can be a laborious task that requires repeated piloting and it is specific to the set of faces the transformations were based on, the exact images, the parameters of the transformations, etc. For this reason, we explore the usefulness of a generic threshold, the minimum distance in a linear image-based space between all 72 of the images used to generate the PCA space.

Gabor

We found a significant negative correlation between the human similarity ratings and the Gabor distances ($r = -.87$, $p = <.001$), shown in Figure 16. The smallest distance between different identities in the Gabor space was 185. Using this threshold, shown as a vertical dotted line in Figure 16, results in 8 points that are misclassified, relative to the human decision boundary at 3.5. As expected, the threshold of 185 is somewhat conservative: a lower threshold of around 160 would produce a very similar classification to the human data, differing only in 3 points. Interestingly, 5 out of 8 of the ‘incorrectly classified’ items are colour transformations. Despite the effect of a change in colour components not crossing the Gabor threshold, the effect of the change is apparently strong enough to decrease perceptual similarity to a point of a ‘dissimilar’ classification. This shows that Gabor distance can be used to quantify PCA transformations and can even be used to predict whether or not a participant is likely to consider two images to depict the same person.

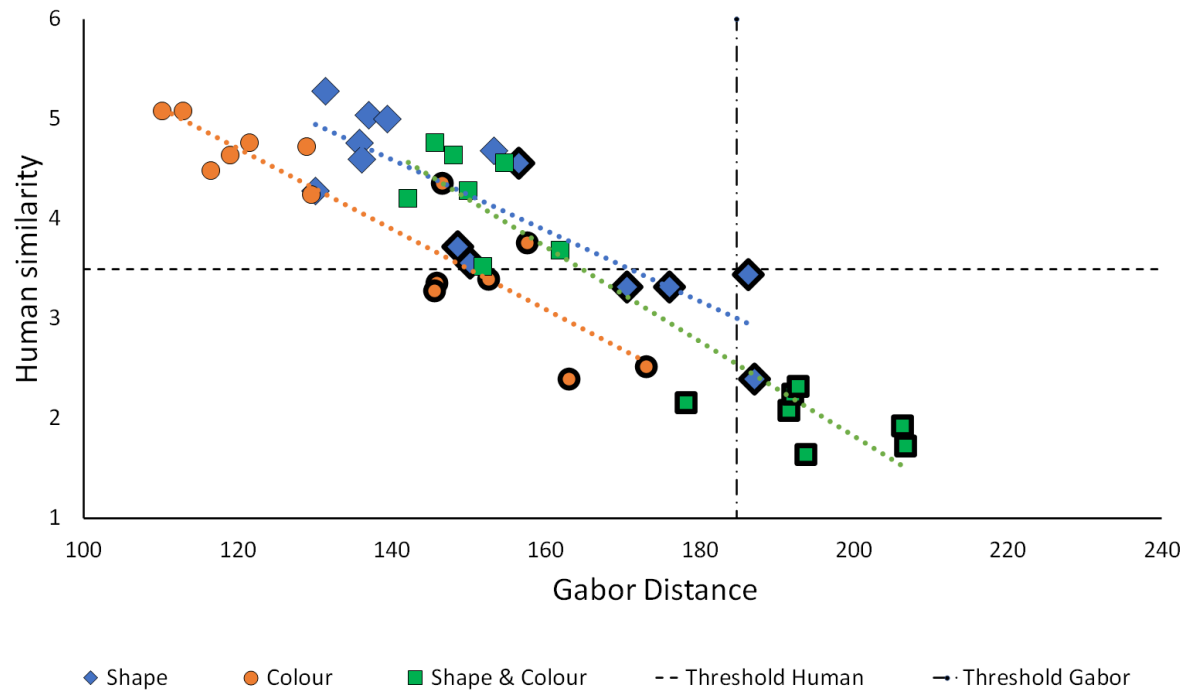


Figure 16 Correlation between perceptual similarity and Gabor distance.

The 6 SD transformations are indicated by a bold outline.

Pixel

We found a significant negative correlation between the human similarity ratings and the Pixel distances ($r = -.85$, $p = <.001$). To avoid non-face image information inflating pixel values, Meytlis (2011) cropped the images to an oval size. In this experiment, a comparable crop has been applied (described above). Using the minimum Pixel distance between the 72 cropped images that make up our PCA space, the threshold is 7350 Pixel distance. The Pixel threshold performed slightly worse than the Gabor threshold, by only correctly identifying 30 out of 42 points (see Figure 17). Again, the colour transformations are the majority of the misclassified points, but only by 1 more error than the transformation of both (colour and shape transformed). Similar to the Gabor threshold and its conservative prediction of dissimilar, the Pixel threshold would benefit from a lower value. For example, lowering the threshold to 4500 would leave only 4 misclassifications.

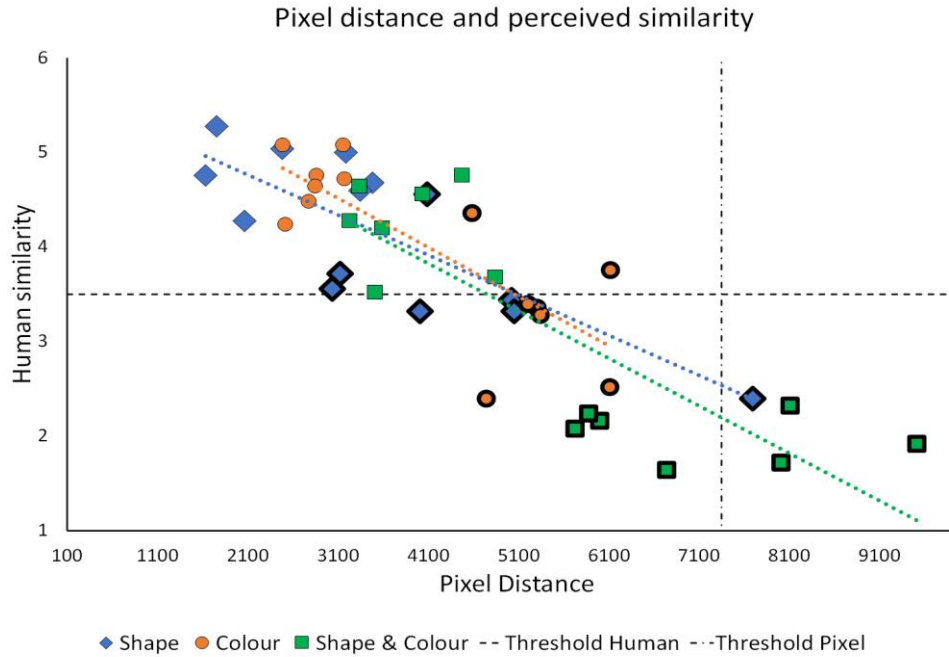


Figure 17 Correlation between perceptual similarity and Pixel distance.

The 6 SD transformations are indicated by a bold outline.

DNN

We found a significant correlation between the human similarity ratings and the DNN distances ($r = .94$, $p = <.001$). This specific DNN uses a value of 0.4 or higher to classify two faces as similar. This threshold provides a slightly less useful guide for predicting perceptual similarity compared to the Gabor threshold and performs the same as the Pixel threshold, as it only correctly identifies 30 out of 42 data points. However, it is important to note that this threshold was not based on our stimulus set but is the threshold used by the DNN. The value of the largest distance between identities in our PCA set (of the average images of 72 actresses) was 0.34, which is close to the threshold set for this DNN. Since the image set in this particular PCA set is relatively small – only 72 – the standard threshold of 0.4 is used.

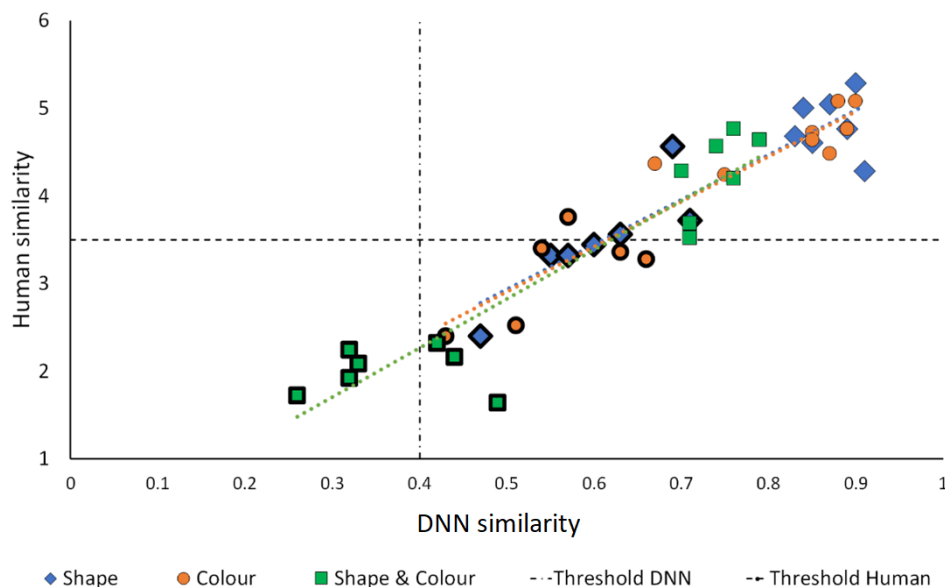


Figure 18 Correlation between perceptual similarity and DNN similarity.

The 6 SD transformations are indicated by a bold outline.

3.5. Discussion Experiment 3

The linear relationship between the DNN similarity space and human similarity is remarkable. The human response scale is, strictly speaking, ordinal: there is no reason to think that the differences between “guess”, “think” and “certain” should be the same number, and the same as the difference between “guess same” and “guess different”. The DNN similarity space is given by the cosine of the angle between the feature vectors. Cosine is clearly non-linear, although most of the non-linearity occurs in the region above 0.8 and most of our points are below that, so within the relatively linear region. Yet these two non-linear spaces, averaged over a number of observations, contrive to give an almost perfect straight-line relationship.

Previous work indicates that humans and DNNs seem to agree on which faces look similar but show less agreement on what makes a face look different (Hancock et al., 2020). One noteworthy aspect of that study was the necessity to use rank correlations, because most of the DNNs appeared to have a marked non-linearity towards the top end of the scale. However, similarity scores of the commercial systems used were not based on the raw similarity between the representational vectors for each pair of images, but on a black-box verification method. It

appears these verification methods have the effect of increasing the reported confidence of a match between sufficiently similar images. We tried some commercial DNNs on the data reported in Figure 6 and they produced a clear ‘hockey stick’ shape (see supplementary analyses). The commercial systems overlook minor differences between face images, to which human observers are sensitive. It therefore seems that the linear relationship observed in the current study results from using the relatively raw similarity between the representational vectors for each pair of images. Thus, understanding the computation of the similarity scores of systems is important when using DNN similarity scores to predict human similarity scores, because the computation affects the form of the relationship.

Similar to the findings of Meytlis (2011) on Euclidean distance, our results show that all three image-based distance values correlate significantly with perceptual similarity. Moreover, we find that a threshold based on the PCA set can be used as a rough prediction of the likelihood that a pair will be considered the same or different identity. Although the results are straightforward, it remains unclear if these linear relationships are also present when multidimensional transformations are used. In Experiment 4a, multidimensional transformations are explored by using transformations with a varying number of principal components changed.

3.6. Experiment 4a: the relationship between the number of principal components transformed, perceptual similarity and familiarity

Background

This experiment is similar to Experiment 3, but with multidimensional instead of single-dimensional transformations. Specifically, we explore the effect of changing different numbers of PCs on perceptual, Gabor, Pixel and DNN space. Based on the findings of Experiment 3, we hypothesize that the relationship between the number of PCs changed and any of the four image spaces is linear. Since it is well known within face identity research that people are better at processing familiar than unfamiliar faces (e.g. Johnston & Edmonds, 2009) and representations of familiar faces are more robust (Burton et al., 2005), we included familiarity to explore the effects of familiarity on perceptual similarity. Therefore, transformations in this experiment were performed on famous faces, as opposed to the unfamiliar, randomly generated faces in Experiment 3. A familiarity check was added in the beginning of the experiment.

Although there is an abundance of research on the effect of familiarity, it remains unclear what underlies this enhanced performance to recognize familiar over unfamiliar faces. A common proposal is that the representation and processing of familiar faces is different from that of unfamiliar faces (R. J. Johnston & Edmonds, 2009; Ramon & Gobbini, 2018). Recent work from Ritchie & Burton (2017) shows that variability is important for learning a new face. Additionally, the variability within images of a specific identity appears to be idiosyncratic information that is key for learning new faces (Burton et al., 2016).

That two images of a person look more similar when familiar is almost the definition of familiarity. Someone familiar with the person can identify them both as being the same individual, while someone unfamiliar has to rely on what they can infer from the appearance (Jenkins et al., 2011). Here, we are applying fixed changes to faces, some of which will be familiar to a given participant and others not. It is plausible that in this situation, viewers familiar with a face will be better able to identify a change: e.g., “that’s X, but the nose is wrong”. They may be able to use their existing memory of the face to identify the changes.

The results from Beale and Keil (1995) make a more subtle prediction. They used morph sequences between different individuals’ faces to test the effect of familiarity on the decision boundary. They found that familiar face pairs had a sharper decision boundary. Moving along the morph continuum, the face looked like the first identity until nearly halfway, when it rapidly shifted to looking like the other identity. For unfamiliar face pairs, the shift was more linear, with a gradual change from one identity to the other. The prediction from this for our imposed changes in appearance is that, for a familiar face, the changed images should look like the original up to some critical point, where it will rather suddenly no longer look like them. If unfamiliar with the original face, then the changes should produce a more linear change in similarity. Therefore, the prediction is that changed faces will initially look more similar when familiar than unfamiliar but that, if the change becomes large enough, this will reverse, with changed familiar faces looking more different.

In these experiments, we are asking participants to decide about probable identity, rather than about similarity directly. So, while those familiar with the face may be able to say, “the nose is wrong”, we expect participants to say “same identity” with more confidence if they are familiar with the face. Whether we see the reversal predicted by the results of Beal and Keil will depend on whether the changes are sufficiently large.

3.7. Methods Experiment 4a

Sample size

Due to the shift to online testing and the novelty of a continuous response scale, with an unknown effect size, it was not possible to set the sample size of the first experiment based on power analysis or simulations. Therefore, sample size is based on a rule of thumb that is commonly used for Linear Mixed Effect Models, which is to have a minimum of 30 data points for 30 individuals per group (Hox, 1998). Familiarity is divided in unfamiliar and familiar, thus there are two groups. Therefore, sample size is 60.

Participants

We tested 61 participants (43 females, 16 males, 1 other, 1 missing value, mean age = 29.9 years, SD = 9, age range = 20 - 68 years. Testing was done through an online platform, Testable. Participants were naive to the purpose of the experiment. All participants agreed to an informed consent form after a written explanation of the procedure. Participation was on a completely voluntary basis without any monetary compensation. The experiment was approved by the local General University Ethics Panel of the University of Stirling.

Apparatus and stimuli

The experiment was programmed through the online platform Testable and participation was only allowed via a PC or laptop. A calibration to control stimulus size was performed by matching the size of an on-screen bar to a credit card, to ensure proper visualisation throughout the experiment. Stimuli preparation and creation were done using Matlab (2019), using the same PCA space as described in Experiment 3. Eight of the 72 average images that went into the PCA were chosen for use in this experiment. The stimuli consisted of these 8 average images, regenerated exactly from the model by using all the components, and variations that had either 2, 3, 5 or 7 colour components (components 10 and 11; 10 – 12; 10 – 14; 10 – 16 respectively) in the PCA space changed. The size of the change was 4 SD. To avoid any directional effects from the PCA transformation, for half the trials the negative transformation and the other half the positive transformation was used. As shown in Experiment 3, any of the three spaces can be used for insight into the size of the perceptual change. We applied this principle in this experiment to ensure that the effect of the transformation was similar for all identities. Note that we would

expect some differences between identities. Suppose a component affects the darkness of the eyebrows. At the image level, a face with thin eyebrows will change less than one with thick eyebrows. Figure 19 shows the Gabor distance between each original image and its four variations. It can be seen that the pattern of differences is very similar for each of the identities, suggesting that we should not expect any marked differences between items in our set. An example of the transformation can be found in Figure 20.



Figure 19 Gabor distance between the original and altered images to check if the transformations had a similar effect in Gabor space for all the identities.

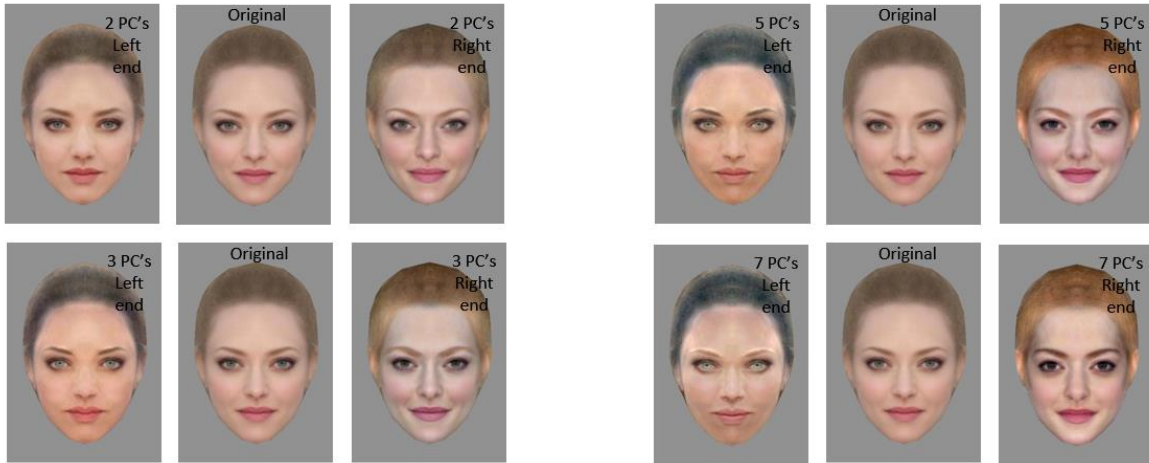


Figure 20 An example of the transformation

The negative end images are -4 SD and the positive end images are +4 SD from the original. The number in the top right of the image indicate how many components are changed. In the experiment the original was paired with transformations of either the negative or the positive end.

Procedure

The experiment consisted of a familiarity check and a face matching task. In the first 8 trials, participants were asked “Would you recognise this person on the street? yes(y)/ no(n)” for each identity (presented in a random order) to establish prior familiarity. Afterwards, participants were presented with the face matching task instructions and had 2 initial practice trials with an identity not included in the actual experiment. This was followed by 32 trials in a random order. In each trial, the original was shown, followed by a mask, the altered image and another mask. Finally, the participants were presented with the question “Could these be two pictures of the same person? 1 – certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. An overview of a trial is shown in Figure 21.

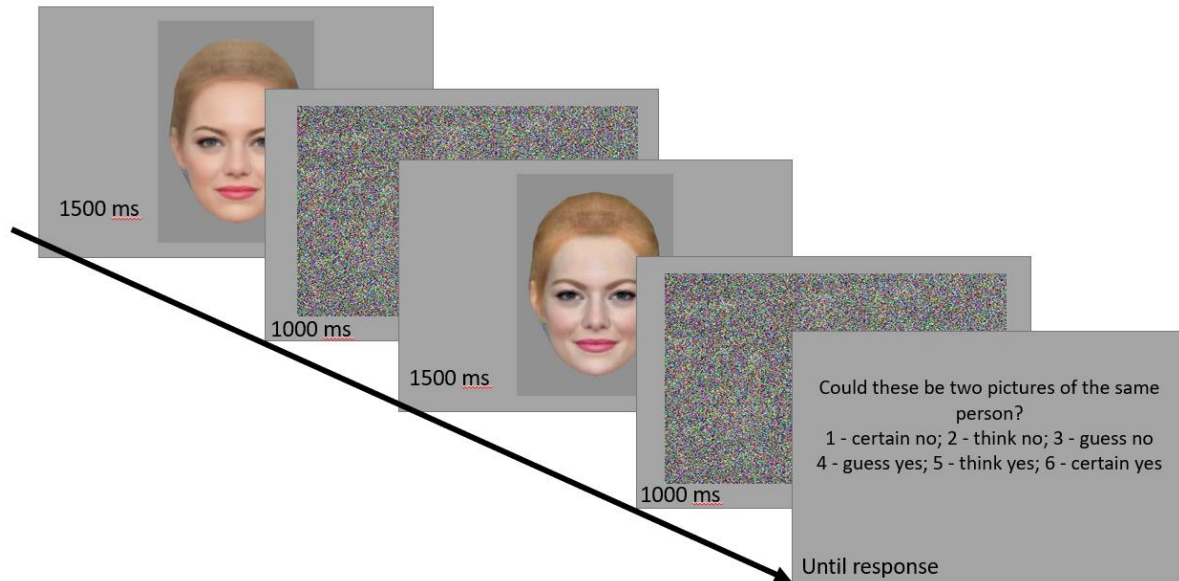


Figure 21 Overview of a trial.

Data analysis

To explore the effects of familiarity, we constructed a linear mixed effects model with the similarity rating as dependent variable. The model included Familiarity and Transformation Level as categorical factors (Familiarity: familiar and unfamiliar rescaled to -0.5 and 0.5 respectively to allow for the analysis of the interaction; Transformation Level: 2, 3, 5, or 7), and the intercept, slope and interaction of familiarity and Transformation Level in the performance per participant as random effects. Due to convergence issues, the main effects were excluded from the random effects. We report the parameter estimate (b), standard error (SE), t value and p value. The linear mixed effect analysis was performed in R (R Core Team, 2019) with *lmerTest* (Kuznetsova et al., 2017). The threshold for significance was set at $\alpha = 0.05$. Normality of the residuals on the model were inspected visually and showed no violation.

3.8. Results and discussion Experiment 4a

In the current study, we explored the effect of multidimensional transformations on the three image-based spaces and perceptual similarity. Similar to Experiment 3, we explore how informative the use of a threshold based on the stimulus set is to predict whether or not the two

images are perceived as the same identity. Lastly, we investigated the effect of familiarity on similarity ratings.

Gabor

We found a significant negative correlation between the human similarity ratings and the Gabor distances ($r = -0.77$, $p < 0.001$). The same threshold as Experiment 3 was used, which was a Gabor value of 185.33 out of 42 points were correctly classified (see Figure 22).

Especially when five components were changed, the estimation was incorrect: 5 out of the 9 incorrect estimations

using the threshold were made in this condition. However, in this experiment only colour components were changed, which were the least accurately predicted in Experiment 3 and this could also explain the misclassifications. In other words, it could be that it is not a certain size of change that is poorly predicted but it might be a poor fit for predicting colour transformations. Similar to Experiment 3, lowering the threshold (to about 167) would increase predictive value.

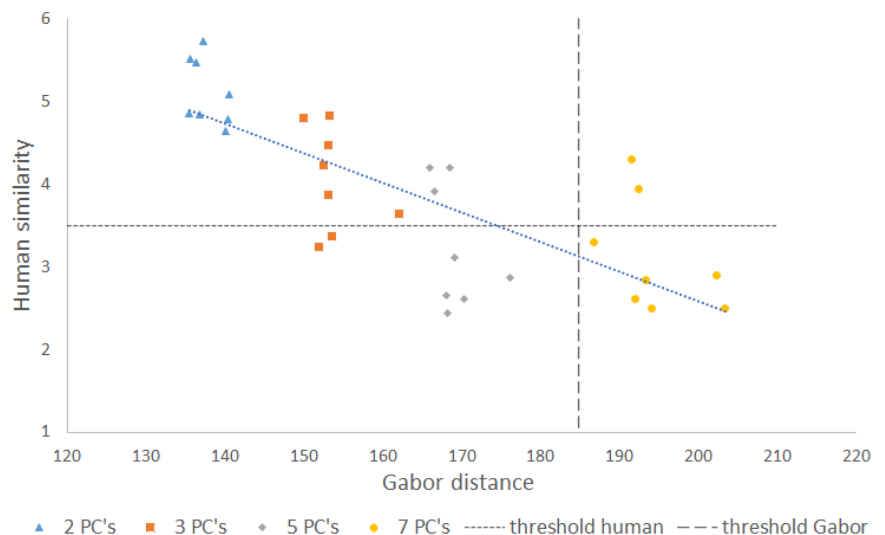


Figure 22 Correlation between perceptual similarity and Gabor distance. The trendline is fitted to the data of all change size groups.

Pixel

We found a significant negative correlation between the human similarity ratings and the Pixel distances ($r = -0.76$, $p = <.001$).

The same threshold as in Experiment 3 was used, which was 7350 Pixel distance. Similar to the Gabor threshold, this provides a decent prediction

whether or not a face pair will be perceived as similar or not

(misclassified points: 33 out of 42; see Figure 23). This accuracy

is similar to that in Experiment 3 (30 out of 42) Again, the estimation using the threshold was poorest in the condition where 5 principal components were changed, whereas the other conditions are quite well estimated. In contrast to the Gabor space, it does not appear that one Pixel threshold can be set to increase accuracy of prediction for the different types of transformations (single- and multidimensional).

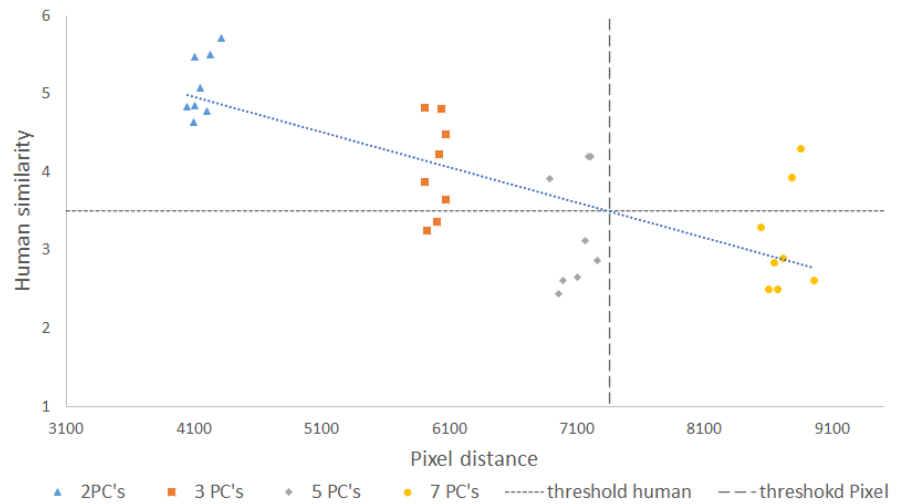


Figure 23 Correlation between perceptual similarity and Pixel distance.

The trendline is fitted to the data of all change size groups.

DNN

We found a significant correlation between the human similarity ratings and DNN distances ($r = .79$, $p = <.001$). As mentioned before, this specific DNN uses a value of 0.4 or higher to classify two faces as similar and a threshold based on our image set was uninformative. Contrary to the first experiment in which the DNN threshold was less useful compared to the other two spaces, this threshold provides an equally useful guide for predicting perceptual similarity by correctly identifying 33 out of 42 data points (see Figure 24).

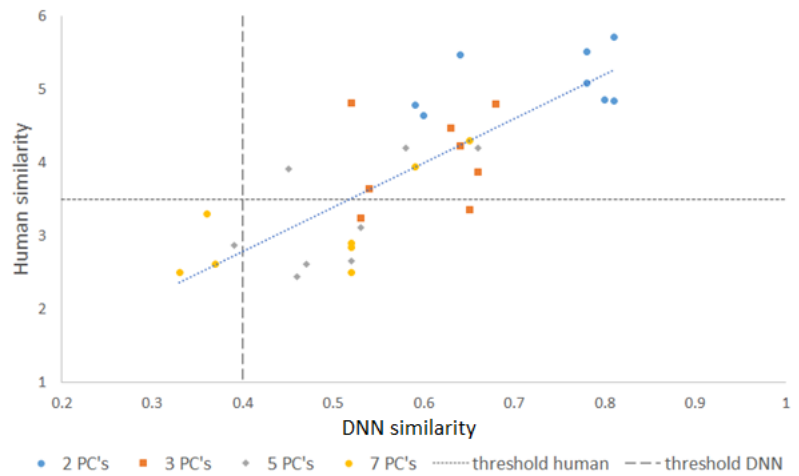


Figure 24 Correlation between DNN distance and perceptual similarity.

The trendline describes all datapoints.

The effect of size of change and familiarity on perceptual similarity

The descriptive statistics of Experiment 4a are shown in Figure 25. Table 8 in the supplementary materials shows the distribution of familiarity of the 8 identities. The results show that increasing the number of components transformed decreased the similarity rating for all levels compared to the base change (2 components changed, all p-values are $p < 0.001$, see Table 9 - Table 11 in Supplementary material – Chapter 3). Interestingly, the similarity rating between 5 and 7 components did not significantly

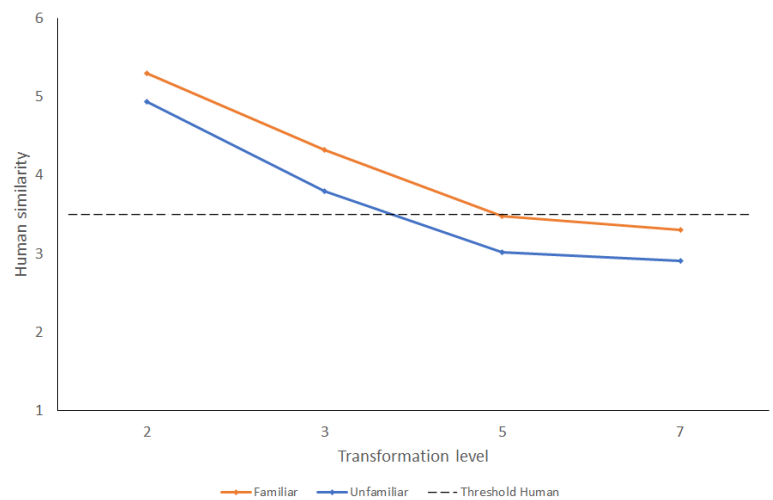


Figure 25 Average similarity rating per Transformation Level grouped by familiarity.

differ (estimate = -0.14, SE = 0.09, $t = -1.62$, $p = 0.106$). Secondly, we found that participants

rated familiar faces as more similar overall (estimate = 0.34, SE = 0.13, $t = 2.59$, $p = 0.010$; see Figure 25). Lastly, there was no interaction found between Familiarity and Transformation Level (see Table 9 - Table 11 in supplementary materials).

The lack of a significant difference in similarity rating between 5 and 7 components (see Figure 25) is surprising because the difference in Gabor space between 5 and 7 components changed is the biggest in the whole stimuli set (see Figure 19). Therefore, one might expect the difference in perceptual distance would be similarly large, but this was not found. As one could see, the relationship is linear until 5 components. This could indicate that Gabor space and perceptual space do not relate linearly for bigger transformations. Thus, the lack of a significant difference could be an indication that more multidimensional transformations are perceived differently. Alternatively, it could be simply an effect of encoding rate and limited presentation duration: the presentation of all faces was 1500 ms so participants might not have had enough time to perceive all the changes.

As described above, the work of Beale and Keil (1995) predicts that changed faces will initially look more similar when familiar than unfamiliar but that, if the change becomes large enough, this will reverse, with changed familiar faces looking more different than when the same transformation is applied to an unfamiliar face. If this prediction became true, there would be a main effect of Familiarity and an interaction effect with transformation level. As expected, we observed greater perceptual similarity for familiar over unfamiliar faces. However, the expected interaction with the Transformation Level was not found. Based on the current experiment alone, it is impossible to say if the prediction with regard to the increase in dissimilarity for familiar faces was wrong because the change could have simply been not great enough to change the identity to the extent needed to increase dissimilarity, i.e., cross and go beyond the border of a category. The current paradigm is an important first step to gaining more insight in how familiarity influences perceptual similarity and it offers a clear-cut way of comparing other transformations that will be used in future research.

3.9. General discussion

In this chapter, we used transformations based on PCA to assess the relationship between perceptual similarity and Gabor, Pixel and DNN distance values. Experiment 3 explored the effects of shape and colour changes on the size of the perceptual change and the three image-based spaces. The results showed that the PCA transformations used in this study result in a linear relationship between the three spaces and human similarity ratings. In addition to assessing the relationship between perceptual similarity and Gabor, Pixel and DNN distance values, Experiment 4a explored the effect of altering different numbers of principal components on the size of the perceptual change, and the effect of familiarity. Similar to Experiment 3, the relationship between each of the spaces and perceptual similarity was linear. Additionally, our findings show that there is a significant relationship between the number of components changed and both the perceptual similarity and the three image-based spaces and that participants rated familiar faces as more similar overall. These findings combined indicate that there is a clear relationship between perceptual and our image-based spaces.

Our findings support those of Yue et al. (2012) that Gabor-based image measures correlate highly with human performance in a discrimination task with complex figures and Meytlis (2011) that there is a relationship between distinctiveness ratings and pixel-wise Euclidean distance. These studies and ours all had different tasks. In Yue et al. (2012), participants performed a match-to-sample task and in Meytlis (2011) participants were asked to rate single faces on distinctiveness and face pairs on perceptual distance (rating how different they looked). In both experiments the stimuli were simultaneously on-screen. Since it is known that availability of stimuli reduces the use of visual working memory and promotes the use of eye movements to gather visual input when needed (Somai et al., 2019), one could argue the observed relationship between perceptual and image-based spaces is due to the ongoing presence of the stimuli that serves as an external representation. Since the external representation is an image and the spaces are image-based, this could mean the findings do not reflect face space but simply the use of visual input. However, in this chapter, Experiment 4a includes a memory component because there is a delay between the two faces rated and the findings are similar. This indicates that the internal representation of faces does indeed relate to image-based spaces.

We do not want to imply that the 3 chosen image-based spaces are superior to any others. This is simply unknown at the time of this thesis, and these spaces were chosen based on

personal experience with them, and the – albeit limited – existing literature (De Valois & De Valois, 1990; Hancock et al., 2020; Jones & Palmer, 1987; Meytlis, 2011; Ringach, 2002; Yue et al., 2012). Additionally, the Gabor and Pixel distance values are easily obtained and understood. The DNN used in this particular instance is not openly available, and other systems with different architecture and training might produce different results (see supplementary material; although the 6 DNNs tested in Hancock, Somai and Mileva (2020) showed similar, significant correlations). In the context of creating a human-like neural network, it will be interesting to compare the differences in DNN spaces and how they relate to perceptual measures.

The primary aim of this chapter was to test the utility of PCA-induced changes in face images to probe human face perception. Our findings show that the graded changes produced by PCA translate to a graded change in perceptual space. In other words, a large PCA change indeed produced a bigger perceptual difference than a smaller one. Additionally, the bigger the number of principal components that were altered, the bigger the perceptual change (up until 5 components). Another argument for the use of PCA-induced changes is the potential insights that can be gained from understanding why it describes perceptual similarity quite accurately in some cases and not in others. For example, in Experiment 4a, 5 out of 8 faces with 5 principal components changed were rated by participants as dissimilar and 3 as similar. One could wonder why this variation exists: what makes some faces appear different with 5 principal components changed while other faces are still perceived the same? Did the specific components changed have a different effect on the different faces? For example, if someone has especially distinctive eyes, then a change to the eye region may have more perceptual effect than on a face with more typical eyes. In the current study, the faces had all the same components changed and no specific selection was made in identities.

To our knowledge, very little research has been done on the role of familiarity in the relationship between (systematic) changes to a face and perceptual similarity. Our secondary aim was to use the paradigm proposed in this chapter to explore this and we found a main effect of familiarity but no interaction effect. Although this work is just the first step in exploring the factors that mediate a change in perceived identity, future work can be easily related to these findings because of the relationship between the image-based spaces. Regardless of the transformations, datasets and images used, it can be quantified and compared.

Our third aim, was the exploration of a threshold based on the face set used to create the PCA space as an aid in stimulus creation. Overall, the three thresholds all gave a decent prediction whether or not a face pair will be perceived as similar or not for each experiment. However, the predictive value of the thresholds varied between the two experiments. The DNN and Pixel threshold made more correct classifications when multi-dimensional rather than single-dimensional transformations were used (single-dimensional: 30/42, multi-dimensional: 33/42), while the Gabor threshold was almost equally informative in both experiments (single-dimensional: 34/42, multi-dimensional: 33/42). The equal performance of the DNN and Pixel threshold might look surprising but is most likely to be consequent of how well-controlled the PCA transformations used are. It is noteworthy that the Gabor threshold was consistent in its mistakes over different types of transformations and could be set lower to increase predictive value, where the Pixel and DNN threshold could not be changed to describe all transformations better. Some might argue that setting a threshold in such a fashion is arbitrary because the threshold would then be different for each stimulus set. However, the goal was never to find the best ‘one-size-fits-all’ threshold, it is merely a suggestion to assist in stimulus creation. Nevertheless, it will be interesting to see how big the variation in the predictive value of the thresholds is between different stimulus sets.

In conclusion, this chapter shows the predictive nature of Gabor, Pixel and DNN distances in the estimation of perceptual changes in faces. This relationship can be used in the exploration of the dimensionality of face space and facilitate in stimulus creation. For example, in research related to variability and learning, the experimenter is often limited in creating a high and low variability set based on observation. Any of the three image spaces can be used to quantify variability and study the effects on a more gradual scale. Lastly, Experiment 4a presents a framework on how to use controlled multidimensional face transformations to explore differences in the perception of familiar and unfamiliar faces. The ability to quantify, and thus control, these transformations is a powerful tool in exploring the factors that mediate a change in perceived identity.

4. Chapter 4 - Exploring other aspects of perceptibility of change: magnitude of change

Chapter main message

In this chapter, I explore whether the pattern found in Experiment 4a would be the same when the changes were smaller in terms of magnitude of change to better understand the relationship between the number of components and magnitude of change, and to explore if the pattern we found in experiment 4a would be the same. Our findings show a similar result as in experiment 4a, but, with higher overall similarity scores, no longer a significant main effect of familiarity and a significant difference between changes with 5 and 7 components. I interpret the meaning of these findings in the context of ‘robust’ representations and face space.

4.1. Introduction

In Experiment 4a, we observed greater perceptual similarity for familiar over unfamiliar faces. One surprising finding, or rather lack thereof, was the lack of a significant difference in similarity rating between 5 and 7 components (see Figure 25), because the difference in Gabor space between 5 and 7 components changed is the biggest in the whole stimuli set (see Figure 19). Since the findings of Experiment 3 and 4a show a linear relationship between perceptual and image-based similarity, one might expect the difference in perceptual distance would be similarly large, but this was not found. So far, I have discussed ‘size of change’ only in the context of the number of components changed. However, it could be possible that other manipulations also affect size of change. More accurately, size of change refers to perceptibility of change, because the facets that are interesting do not only exist in image-based spaces but affect perceptual space. In this chapter, I will explore possible reasons for the findings in Experiment 3 and 4a by replicating the experiment with another facet of perceptibility of change, magnitude of change, by altering the magnitude of changes as well as the number of PCs changed.

One plausible reason I proposed in the last chapter is that the lack of a significant difference in similarity rating between 5 and 7 components indicates Gabor space and perceptual space do not relate linearly for bigger transformations. However, bigger transformations so far have only meant a higher number of principal components changed. The magnitude of change

used in Experiment 4a was 4 standard deviations, to ensure that the size was large enough to change identity, in other words, cross the identity border. The smaller changes, of 2 and 3 components changed, are changes within the same identity category according to the similarity rating. The larger changes, of 5 and 7 components changed, are the changes that crossed the identity border and are transformations that both fall outside the identity boundary. This contrast between the smaller and larger changes, the transformations being within identity or in this space outside of the identity boundary should be considered as a reason for the lack of a significant difference in similarity rating between 5 and 7 components since it is the most obvious difference between the components. However, if the identity border matters, it seems odd that the unfamiliar faces followed a similar pattern as the familiar faces, as the work of Beale and Keil (1995) would predict lower similarity rating for familiar face once they cross the identity border compared to unfamiliar faces. Based on the differences processing familiar and unfamiliar faces in the Beale and Keil (1995) study, you would expect to see some interaction for changes that cross the identity border.

To explore if the pattern remains the same if the changes do not cross the identity border, I repeat the experiment with a smaller magnitude of change of 2 standard deviations. Additionally, I also included a 1 SD version of the 7-component change, to explore if participants can detect those changes. If the changes of the smallest magnitude are perceived, the effect of familiarity should be present. If crossing the identity border influences the relationship between perceptibility of change and familiarity with perceptual similarity, the findings of Experiment 4b should differ from Experiment 4a, although it is difficult to specify what the resulting pattern would be. However, an indication of the identity border as a relevant factor for the findings in Experiment 4a would be a significant difference in similarity rating between 5 and 7 components, since this was not present in Experiment 4a. A secondary goal of this experiment is continuing the search for well-controlled high dimensional stimuli suitable to explore any potential differences related to familiarity.

As another alternative explanation for the lack of a significant difference in similarity rating between the changes based on 5 and 7 components, I argued it could be simply an effect of encoding rate and limited presentation duration: the presentation of all faces was 1500 ms so participants might not have had enough time to perceive all the changes in the 7 components changed condition. If an encoding rate capacity is the cause, changing the magnitude of the

change should not change the pattern of the findings of this experiment compared to Experiment 4a.

Predicting the effect of a smaller change in terms of magnitude is difficult, because the size of the change still only reflects variation in our specific stimulus set and it remains uncertain how it relates to general variation in a complete face diet. Despite the standard deviation feeling intuitively interpretable, it is important to keep in mind that the standard deviation only reflects the variation within the PCA set used for the PCA. Since the set used is based on 72 famous, mostly white, averages of actresses, this leads to a naturally smaller variation than a set that is more inclusive. On the one hand, it could be possible that the relatively small variation in this set might decrease the effect of magnitude of change and make the number of components changed matter more. On the other hand, the lower variation resulting from our well-controlled images can lead to clearly defined components, especially the earlier components that were changed. The clear components could then have a great perceptual effect. Thus, the relationship between number of components changed and magnitude of change needs further study.

In short, the aim of the following experiment is two-fold. Firstly, the experiment is to better understand the relationship between the number of components and magnitude of change. Secondly, it is to explore if the pattern we found in experiment 4a would be the same when the changes were smaller in terms of magnitude of change and does not cross the identity-border. For these reasons, the experiment was re-run with smaller changes of 2 standard deviations. Additionally, we've made the changes more natural /average looking to increase difficulty and gain a greater control of the transformations.

4.2. Methods Experiment 4b

Participants

Since this experiment is a variation on Experiment 4a and the aim is to interpret the results across experiments, the sample size was matched to Experiment 4a. We tested 60 participants (34 female, 24 male, 2 other, mean age = 32.5 years, SD = 10.5 age range = 18 - 57 years. Testing was done through an online platform, Testable, recruitment was done through Prolific. Participants were naive to the purpose of the experiment. All participants agreed to an informed consent form after a written explanation of the procedure. Participation received monetary compensation. The

experiment was approved by the local General University Ethics Panel of the University of Stirling.

Apparatus and stimuli

Similar apparatus and stimuli as in Experiment 4a (Chapter 0) was used, with a small adjustment to the stimulus creation. In Experiment 4a we observed several of the faces with the larger transformations to become more distinctive relative to other faces. To reduce a strategy based on how natural or average the face looked, the way the transformations are applied was adjusted. In the previous experiment, a standard deviation of 4 was used and the change was absolute: the changes were simply a certain distance added to the value of a particular component. However, different faces have different base values and faces that were already far from the average that had even more added to them could become more distinctive but also more unnatural than a change in the other direction would cause. To avoid any strategy utilizing a more distinctive look, the direction of the transformation was changed so that it will always cross the value of the average of that component. Specifically, for each component if the value was positive it was swapped to negative. Subsequently, the same transformations as in Experiment 4a were used. This correction causes the negative end of the change to become more extreme and the positive end shift toward the average of the stimulus set (for an illustration of a 5-step change, see Figure 26). For this reason, solely the positive transformation was used in this experiment.

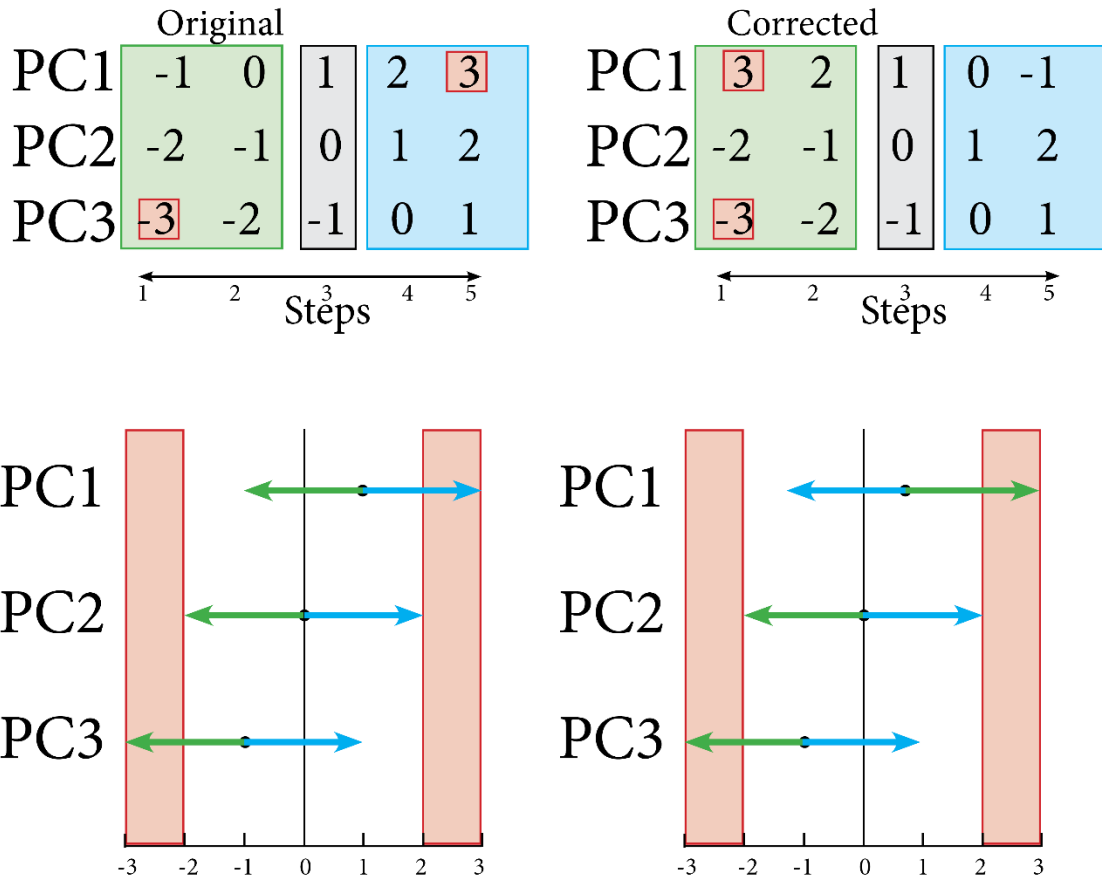


Figure 26 Schematic visualization of the image transformation for the stimuli creation of Experiment 4a (left) and Experiment 4b (right, after correcting for extreme values).

The tables represent the value of the components when a change of 1 SD is made and $SD = 2$. For the sake of simplicity, all components have the same SD, but the actual code calculates the SD per component. The new values are calculated for all the steps and the middle step always represents the original. The graphs depict the transformations, where the green arrows are the values of step 1 and 2, and the blue arrows are the values of step 4 and 5. In the original stimulus creation step 1 and 2 would be the negative steps, and 4 and 5 the positive steps. To avoid the effects of extreme values of components, the direction of the transformation was changed so that it will always cross the value of the average of that component. This can be seen in PC1 in the top row of the tables: the value of step 5 is flipped with step 1, and the value of step 4 is flipped with step 2. This ensures that the last steps are less extreme, i.e. more towards the average.

Stimuli preparation and creation were done using Matlab (2020), using the same PCA space as described in Experiment 3. The same 8 identities as in Experiment 4a were used. The stimuli consisted of these 8 average images and variations that had either 2, 3, 5 or 7 colour components (components 10 and 11; 10 – 12; 10 – 14; 10 – 16 respectively) in the calculated PCA space changed, but now corrected so the change would cross the average value for each of the changed PCs. The size of the change was 2 SD. Additionally, participants also saw the 7-

component change at a magnitude of 1 SD to explore if they could perceive those changes. The 1 SD-changes are analysed separately. An example of the transformation can be found in Figure 27 and an overview of a trial in Figure 28. The descriptive statistics of experiment 4b are shown in Figure 29. The procedure and analysis remained the same as in Experiment 4a.

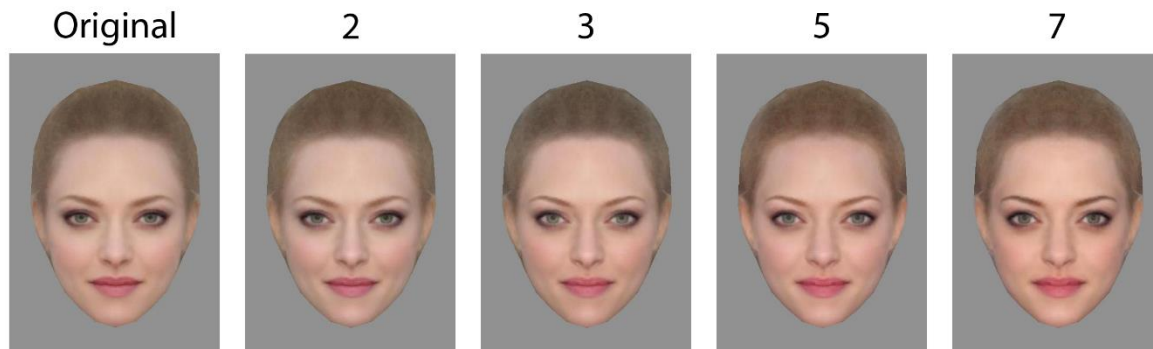


Figure 27 Example of the PCA transformation using 2 SD and a change towards the average.

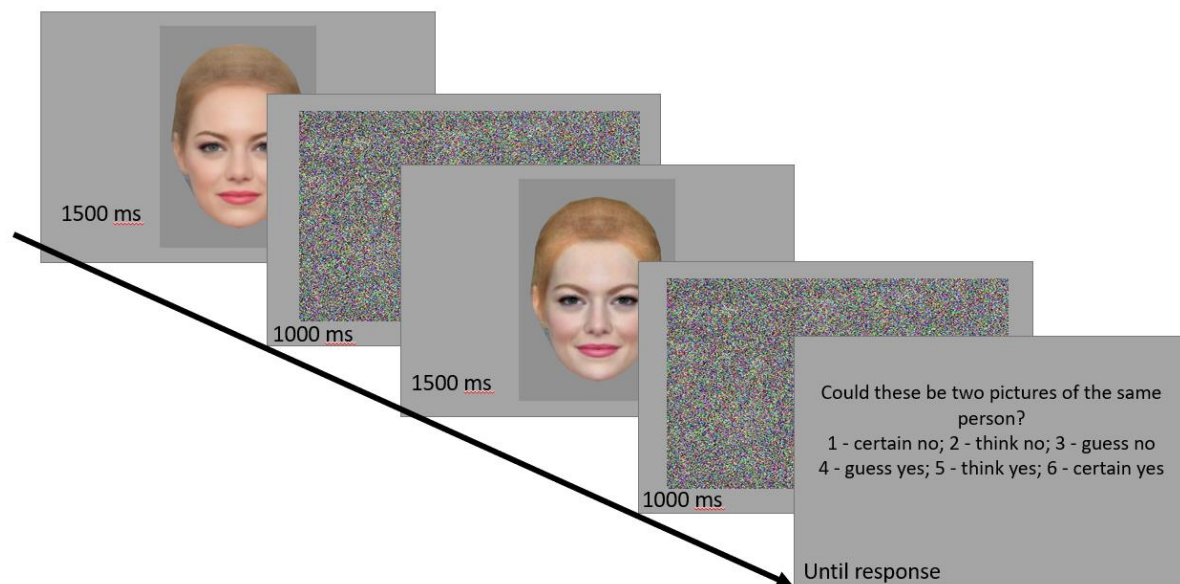


Figure 28 Overview of a trial

Data analysis

To explore the effects of familiarity, I constructed a linear mixed effects model with the similarity rating as dependent variable. The model included Familiarity and Transformation Level as categorical factors (Familiarity: familiar and unfamiliar rescaled to -0.5 and 0.5 respectively to allow for the analysis of the interaction; Transformation Level: 2, 3, 5, or 7), and the intercept, slope and interaction of familiarity and Transformation Level in the performance

per participant as random effects. Due to convergence issues, the main effects were excluded from the random effects. To explore the main effect of Familiarity, a similar model without the interaction effects included was used. I report the parameter estimate (b), standard error (SE), t value and p value. The linear mixed effect analysis was performed in R (R Core Team, 2019) with *lmerTest* (Kuznetsova et al., 2017). The threshold for significance was set at $\alpha = 0.05$. Normality of the residuals on the model were inspected visually and showed no violation.

4.3. Results Experiment 4b

The descriptive statistics of Experiment 4b are shown in Figure 29. The results show a similar pattern as in Experiment 4a. There is a significant main effect of familiarity (estimate = 0.26, SE = 0.08, $t = 3.46$, $p = 0.077$).

Even with a smaller change, all transformation levels significantly differed from the

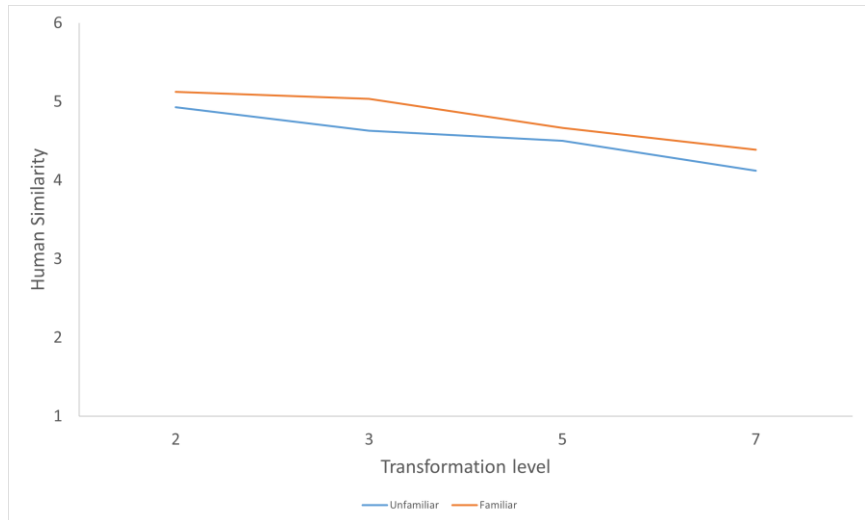


Figure 29 Average similarity rating per Transformation Level grouped by familiarity.

base level (2 components changed). Additionally, there was no significant interaction effect between Familiarity and Transformation Level. Contrary to Experiment 4a, even the difference between Transformation Level 5 and 7 was significant in this experiment (component 5 used as base, estimate = -0.33, SE = 0.09, $t = -3.68$, $p < 0.001$). To provide a clearer comparison between Experiment 4a and b, a table with the descriptives of both experiments is shown in Table 1.

Table 1. Average similarity rating per Transformation Level.

	2	3	5	7
Experiment 4a	5.11	4.06	3.25	3.11
Experiment 4b	5.03	4.83	4.58	4.26

Exploring detectability with 1 SD changes

Apart from the 2 SD change, participants also saw the 7-component change at a magnitude of 1 SD. To explore the effects of familiarity, I constructed a linear mixed effects model with the similarity rating as dependent variable and included Familiarity as a categorical factor, and the intercept and slope per participant as a random effect. The results show a significant effect of familiarity (estimate = 0.23, SE = 0.11, $t = 2.01$, $p = 0.04$).

4.4. Discussion Experiment 4b

In this experiment, I explored if the pattern found in Experiment 4a would be the same when the changes were smaller in terms of magnitude of change but the same in terms of number of components changed. This experiment had two aims: to better understand the relationship between the number of components and magnitude of change, and to explore if the pattern we found in experiment 4a would be the same when the changes were smaller in terms of magnitude of change and does not cross the identity-border. Our findings show a similar result as in experiment 4a, but with higher overall similarity scores. Additionally, the difference between changes with 5 and 7 components changed was not significant in Experiment 4a but is significant here.

One noticeable difference between Experiment 4a and 4b is that the similarity ratings in Experiment 4b remain higher than in 4a, except for the smallest transformation. Although this might seem like an obvious result, it is actually quite an important one. As mentioned before, the PCA set used throughout this thesis is based on 72 famous, mostly white, averages of actresses and this leads to a naturally smaller variation than a set that is more inclusive. In other words, if we could calculate the PCA of all the faces in someone's face diet, a standard deviation of change would lead to greater perceptual changes. Even with the current small effects of standard deviation, it was enough to reduce the size of the change to stay within the identity borders. Thus, it appears that magnitude of the change is an important factor of size of change.

As mentioned before, in the study of Beale and Keil (1995), the identity scale explored was from 1 identity to another, thus the crossing of a border instantly led to another identity. In the current study, it just moves away from the identity. The significant Familiarity effect in the 1 SD condition was therefore especially surprising: That condition was originally only put in to confirm that such a small change would not have an effect, but then it did. Thus, even with a

small change in terms of magnitude, the number of components changed matter. This result emphasizes the running theme in this discussion: the need for a clear and complete operational definition of perceptibility of change.

Lastly, there was a significant difference between the changes with 5 components changed and with 7 components changed, that was not present in Experiment 4a. At some level, it is difficult to compare the results of Experiment 4a and 4b to each other, as the new procedure to create the stimuli mostly affected the larger changes. Thus, the difference in findings between the two experiments could simply be because the changes looked more similar with the new procedure, which caused less perceptual disturbance and made it easier to distinguish the two largest transformations. Alternatively, it could be a meaningful finding, where changes after the decision boundary are just more difficult to distinguish. To address such a question and the questions raised throughout this discussion, it is of obvious importance to get a better understanding of the factors that can influence perceptibility of change in order to control the relevant image properties and properly assess the effect of crossing the identity border.

These results taken together with experiment 4a indicate that the size of the change is not only influenced by the number of components, but also the magnitude of the change. Although the significant difference in similarity rating between 5 and 7 components is an indication of the identity border influencing the relationship between perceptibility of change and familiarity with perceptual similarity, more studies on the different aspects of perceptibility of change are needed in order to make more conclusive statements. Untangling the different facets of perceptibility of change will bring us closer to understanding what changes matter, what makes a face distinctive and, in general, will provide a useful framework for the exploration of face space. In the next experiment, I will look at one more potential facet of perceptibility of change, the selection of principal components.

5. Chapter 5 - Title: Familiarity and quantifying sensitivity to dissimilarity

Chapter main message

In the previous chapters, we (Somai & Hancock, 2022) found that PCA could be used to create well-controlled, scaled transformations that provide a controlled way to alter the appearance of a face. In the current study, I aim to use those specific PCA transformations to quantify sensitivity to change. A 2-alternative forced-choice task was used to explore differences between familiar and unfamiliar faces in the ability to identify changes made to the appearance of the face. The task was carried out with transformation of 2SD (Experiment 5a) and 4 SD (Experiment 5b) changes. Participants were asked to select the face that more closely resembled the identity presented at the start of the trial. As expected, I found a main effect of perceptibility of change (in number of components) between the response pairs. Additionally, I found an interaction between Familiarity and perceptibility of change. This means that the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces. Both experiments revealed a similar pattern of results which indicates a potential memory effect, for which I discuss several reasons.

5.1. Introduction

In Chapter 0, we found that familiar faces are perceived as more similar overall compared to unfamiliar faces (Somai & Hancock, 2022). In those experiments (Experiment 4a) a transformed image was compared to an image with the original identity. Although this finding seems sensible and is in line with the idea of a more robust representation for familiar faces compared to unfamiliar ones (Burton et al., 2005; Kramer et al., 2018), these results provide little further insight into what ‘robust’ means. It can even feel counter-intuitive if you try to interpret the findings in terms of sensitivity to change. ‘Robust’ is often interpreted as a richer internal representation, i.e. containing more information for a familiar faces compared to an unfamiliar face. If that were true, you would expect that the same transformation for a familiar and unfamiliar face should be more obvious in the familiar face, as there is more detailed information present to compare the change to. However, this does not appear the case and that might be because the argument above is formulated based on a task that focuses on sensitivity to similarity. In this chapter, I further explore the relationship between familiarity and perceptual similarity by shifting our focus from similarity to sensitivity to change (i.e., dissimilarity).

Although in a completely different framework, some predictions can be made on the basis of the work of Menon et al. (2015) in which the authors aimed to determine whether it is specifically the identity-level representation of a face that mediates unfamiliar face processing as opposed to a pictorial representation. In this study, participants were presented with an image and asked if these matched their target. Menon et al. (2015) manipulated the identity-related information by telling participants they were either looking for 1 identity depicted by 2 images or 2 identities depicted by 1 image each, without there actually being a second identity present. This design allowed for the image information to be kept the same in both conditions. Participants who noticed the misinformation were excluded (if they recognized the 2 images depicted 1 person in the condition that they were told it was 2 different people). Participants were instructed to look at the target images until they felt they could remember them. Menon et al. (2015) found that identity-level information indeed mediated unfamiliar face matching, but only when the task was presented sequentially opposed to simultaneously. The authors suggested the sequential presentation shifted the image-based strategy to one emphasizing identity and is necessary for effects associated with the identity-level representation.

In the previous chapter, we explored the effect of familiarity on the relationship between well-controlled image transformations and perceptual similarity. Here, I shift the focus from identity categorization to sensitivity to change. To explore differences between familiar and unfamiliar faces in the ability to perceive the different levels of change, a 2 alternative forced choice (2-AFC) paradigm is used. In this task, two altered images are compared to an original and the participant is asked to pick the image that resembles the original more closely. The specific PCA transformations used in the previous studies are again used to quantify the different levels of change. As mentioned above, Menon et al (2015) suggested that sequential presentation shifted the image-based strategy to one emphasizing identity and is necessary for effects associated with the identity-level representation. In other words, they suggest that their effects were dependent on the use of internal representations in the task. Based on the discussed work of (Menon et al., 2015) and the presented work so far, the effect of familiarity should become evident with a sequential task design, which leads to the following hypothesis: if familiar faces have a richer internal representation compared to unfamiliar faces, accuracy to select the most similar looking transformation should be higher than for unfamiliar faces.

Two different magnitudes of change are used: 2 and 4 SD. The 2 SD changes are used to explore the sensitivity to change within the same identity and the 4 SD changes for across-identity boundaries.

5.2. Methods Experiment 5a: sensitivity to change within identity (2 SD change)

Sample size

Similar to Experiment 4a, sample size is based on a rule of thumb that is commonly used for Linear Mixed Effect Models, which is to have a minimum of 30 data points for 30 individuals per group (Hox, 1998). Familiarity is divided in unfamiliar and familiar, thus there are two groups. Therefore, sample size is 60.

Participants

58 participants (44 female, 14 male, mean age = 33.8 years, SD = 11.0 age range = 18 - 65 years) were tested. Testing was done through an online platform, Testable, recruitment was done

through Prolific. Participants were naive to the purpose of the experiment. All participants gave consent form after reading an explanation of the procedure. Participation received monetary compensation. The experiment was approved by the General University Ethics Panel of the University of Stirling.

Stimuli

The experiment was programmed through the online platform Testable and participation was only allowed via a PC or laptop. A calibration to control stimulus size was performed in form of matching the size of an on-screen bar to a credit card, to ensure proper visualisation throughout the experiment. The stimuli creation followed a similar procedure as used in Experiment 4b (Chapter 0), which changed the components in the direction of the average (Figure 30) to ensure more natural looking changes.

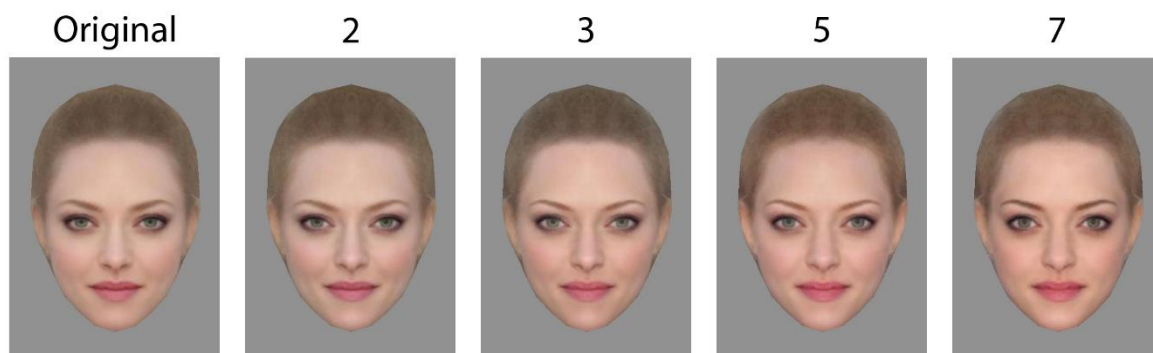


Figure 30 Example of the transformations

Procedure

The experiment consisted of a familiarity check and a face matching task. The same familiarity check as previous experiments was used: participants were asked “Would you recognise this person on the street? yes(y)/ no(n)” for each identity (presented in a random order). Afterwards, participants were presented with the face matching task instructions and had 6 initial practice trials with two identities not included in the actual experiment. This was followed by 36 trials in a random order. In each trial, the original was shown, followed by a mask, the first altered image, another mask, the second altered image and a last mask. Finally, the participants were presented with the question “Which of the two changed pictures more closely resembles the person shown in the first image 1 - first picture; 2 - second picture”. All transformation levels of the previous experiment were used, and all possible combinations between image pairs were used for all

participants (an image with 2 components changed paired with an image with 3 components changed, 2-5, 2-7, 3-5, 3-7 and 5-7). The experiment was counterbalanced on which Transformation Level was shown first. An overview of a trial is shown in Figure 31.

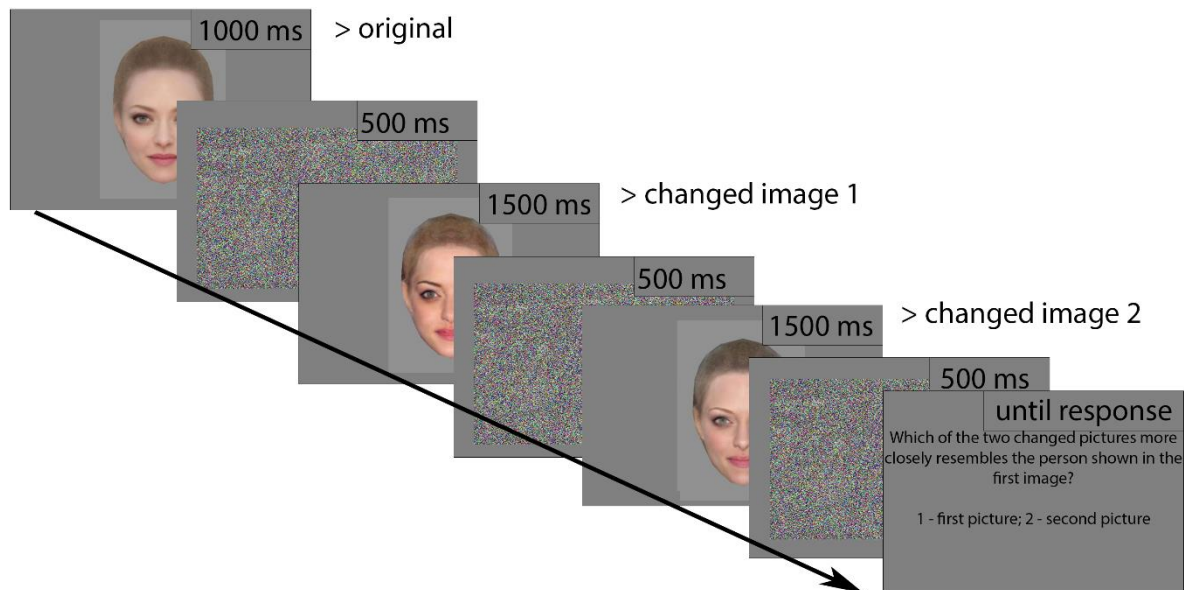


Figure 31 Example of a trial

Data analysis

To explore the effects of familiarity on discrimination ability, I constructed a logistic model with the correct response as dependent variable. To analyse the distance between the pairs, the different combinations of image pairs were recoded into 'pair distance', which was the difference between the number of components altered in the two images. Two combinations fell under category 2 (3-5 and 5-7), and all other categories were made up of single combinations. The model included Familiarity and Pair Distance as categorical factors (Familiarity: familiar and unfamiliar coded as -0.5 and 0.5 respectively to allow for the analysis of the interaction; Pair Distance: 1, 2, 3, 4, or 5), and the intercept, slope and interaction of Familiarity and Pair Distance in the performance per participant as random effects. To explore the main effect of Familiarity, a similar model without the interaction effects included was used.

The parameter estimate (b), standard error (SE), z-value and p-value is reported. The logistic model was performed in R (R Core Team, 2019). The threshold for significance was set at $\alpha = 0.05$.

Table 2 components per pair distance

Pair Distance	Components changed in pair of altered images
1	2-3
2	3-5; 5-7
3	2-5
4	3-7
5	2-7

5.3. Results Experiment 5a

The descriptive statistics of Experiment 5a are shown in Figure 32. The expected main effect of pair distance is found, but only for Pair Distance 4 (reference Pair Distance is 1; estimate = 0.44, SE = 0.16, $z = 2.67$, $p = 0.008$) and Pair Distance 5 (estimate = 0.59, SE = 0.17, $z = 3.53$, $p < 0.001$). There is a significant main effect of Familiarity (estimate = 0.26, SE = .09, $z = 2.74$, $p = .006$). Additionally, there is an interaction with Transformation Level at Pair Distance 5 (reference Pair Distance is 1 and Familiarity is ‘unfamiliar’; estimate = 0.68, SE = 0.34, $z = 2.04$, $p = 0.041$).

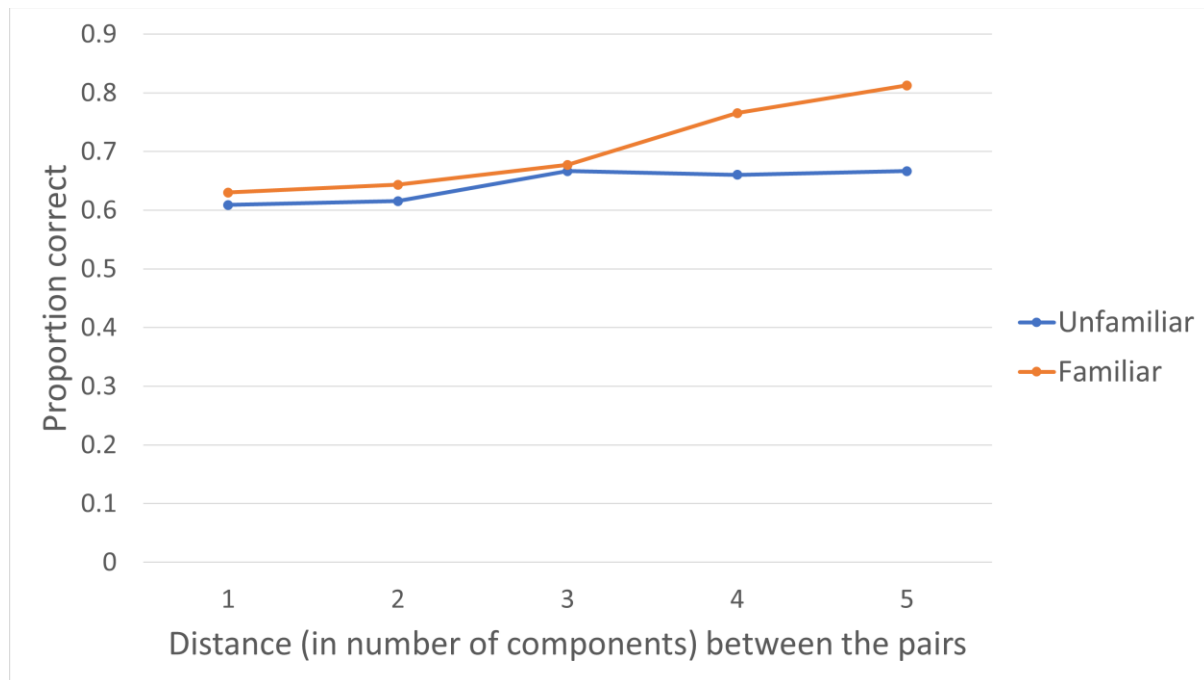


Figure 32. Proportion correct per pair distance.

5.4. Discussion Experiment 5a

These results are in line with the result found in Experiment 4b in which the relationship between perceived similarity and transformation level of a magnitude of 2 SD was explored. To briefly recap, in Experiment 4b two faces were presented sequentially to the participant along with the question “Could these be two pictures of the same person? 1 – certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. Contrary to Experiment 4a and 4b, the current experiment does show a significant interaction between Pair distance and Familiarity at the largest distance between image pairs.

Since I only tested up to a pair distance of 5 and the transformation size was 2 SD, the interaction effect is difficult to interpret because the identity boundary is likely not crossed. To clarify, a pair distance of 5 means that the participant, after seeing the original, saw 1 altered image transformed with 2 components and 1 with 7 components changed. In Experiment 4b, an image with 7 components changed is still perceived well within the identity boundary (if compared to the original): the average similarity rating given was 4.26 (4.12 and 4.39 for unfamiliar and familiar faces respectively). Anything higher than 3.5 is considered the same identity. An image with 2 components changed compared to the original had an average similarity rating of 5.03 (4.93 and 5.12 for unfamiliar and familiar faces respectively). Thus, even at the largest Pair Distance tested in Experiment 4a, the identities are most likely perceived as the same identity. As discussed in Chapter 0, the work of Beale and Keil (1995) found that familiar face pairs had a sharper decision boundary the prediction than unfamiliar faces. This means that familiar faces will initially look more similar, but if the change becomes large enough it will look more different, while this happens more gradually for unfamiliar faces. Therefore, it is unclear if the interaction effect occurs at all large differences between the two shown images if the Pair distance is greater than 3, or if the pattern of results is different for Pair distances greater than 5, because the decision boundary is not crossed in this experiment.

Discussing the role of the identity boundary in the relationship between familiarity and sensitivity to change is highly speculative at this point and it is impossible to draw any conclusions based on the data there is available from the current experiment. However, if the identity boundary plays a role in the interaction effect between Pair Distance and Familiarity, a bigger change magnitude – one that changes perceived identity (such as used in Experiment 4a) - should change the pattern of results. That possibility is explored in Experiment 5b.

5.5. Experiment 5b: crossing the identity boundaries (4 SD change)

Introduction

In Experiment 5a, the magnitude of the transformation was 2 SD. From Experiment 4b, we know that the similarity ratings between an original image and an image transformed with components in the magnitude of 2 SD are still considered as the same identity. In the current experiment, I explored the sensitivity to change when the transformations' size in terms of SD are large enough to change identity. Specifically, I will use the same transformations but with a magnitude of 4 SD.

The work of Beale and Keil (1995) provide us with some indication on how the results of Experiment 4a might change if a larger transformation is used. They found that familiar face pairs had a sharper decision boundary the prediction than unfamiliar faces. In the current experiment, this would mean that once the Pair Distance is large enough for 1 of the altered images to cross the identity-boundary, the perceived difference between the two altered images should be larger for familiar faces than for unfamiliar faces.

To gain insight into what Pair Distances are cross-boundary, the similarity ratings of Experiment 4a can be used, in which two faces – one original, one altered with changes of 4 SD – were presented sequentially to the participant along with the question “Could these be two pictures of the same person? 1 –certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. An overview of the average similarity ratings of Experiment 4a are given in Table 3. Anything higher than 3.5 is considered the same identity. Overall, based on the similarity ratings, changing 5 and 7 components changes the perceived identity. Changing 3 components appears to shift faces closer to the identity-boundary when unfamiliar faces (rated 3.79) than familiar faces (4.32). As the Pair Distance is the difference in number of components changed between the altered images, the Pair Distances that are cross-identity are 3 (a pair of altered images with 2 and 5 components changed) and 5 (a pair of altered images with 2 and 7 components changed).

The findings of Experiment 5a, specifically the interaction effect, indicates that the effect of Familiarity is visible when the Pair Distance is large enough, so I would expect to also find a difference in Pair Distance 4 in the current Experiment. The increased size of the transformation,

4SD instead of 2SD should increase difference between the pairs of images. Thus, due to the cross-identity pairs and larger change, I hypothesize that there will be a main effect of Familiarity, in addition to an interaction between Familiarity and Pair Distance.

Table 3. Averages per number of components changed.

Components changed	Average similarity rating	Average similarity rating Familiar faces	Average similarity rating unfamiliar faces
2	5.11	5.29	4.93
3	4.06	4.32	3.79
5	3.25	3.48	3.02
7	3.11	3.30	2.91

Methods Experiment 5b

The methods are exactly the same as in Experiment 5a, except that the transformations on the images were 4 SD opposed to 2 SD. These stimuli are similar as the ones used in Experiment 3a but were created with the procedure described in Experiment 4b (Figure 33). As mentioned in chapter 0, this was done to make more natural images that always moved a face more towards the average. Since this experiment is a variation on Experiment 5a and the aim is to interpret the results across experiments, the sample size was matched to Experiment 5a. 60 participants (35 female, 25 male, mean age = 30.3 years, SD = 10.5 age range = 18 - 62 years) were tested.

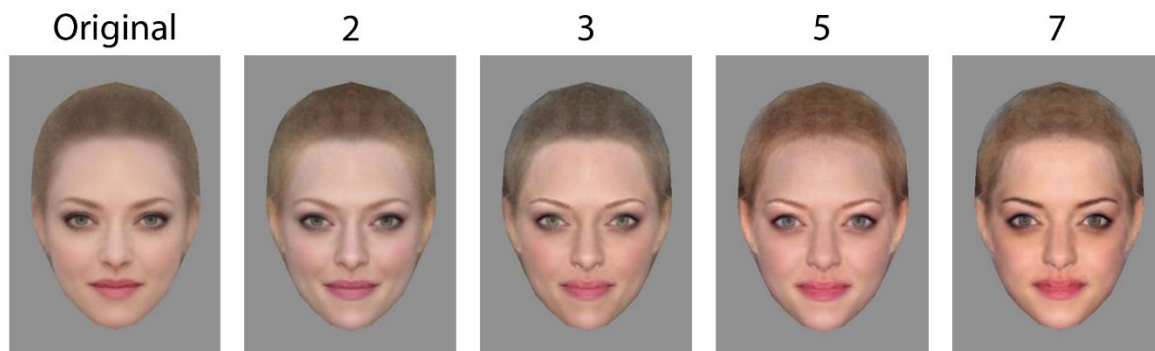


Figure 33 Example of the transformations.

5.6. Results Experiment 5b

The descriptive statistics of Experiment 5b are shown in Figure 34. The expected main effect of Pair Distance is found for all Pair Distances. Interestingly, performance for Pair Distance 2 is significantly worse than Pair Distance 1 (estimate = -.40, SE = 0.15, $z = -2.69$, $p = .007$). The

results show that Familiarity has a significant main effect on the sensitivity to change (estimate = .28, SE = 0.11, $z = 2.55$, $p = .011$). Again, Familiarity does have a significant interaction with Pair Distance if the Pair Distance is 5 (reference Pair Distance is 1 and Familiarity is ‘unfamiliar’; estimate = 1.10, SE = 0.49, $z = 2.24$, $p = .025$).

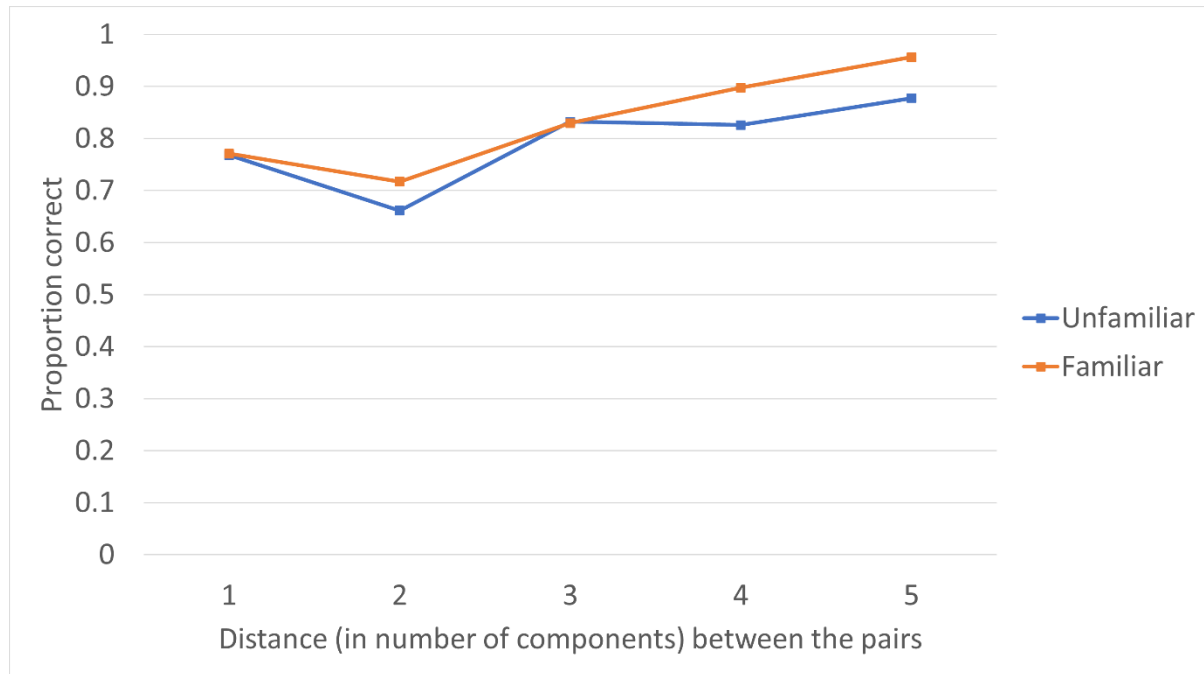


Figure 34. Proportion correct per pair distance

5.7. Discussion Experiment 5b

In this experiment, I explored if the identity-boundary plays a role in the interaction effect between Pair Distance and Familiarity by using a bigger change magnitude – one that changes perceived identity. A significant main effect of Pair Distance, Familiarity, and interaction between Familiarity and Pair Distance was found.

Due to the cross-identity pairs and larger change, I hypothesized that there will be a main effect of Familiarity, in addition to an interaction between Familiarity and Pair Distance.

The main effects found for Pair Distance and Familiarity are in line with Experiment 4a and 4b. The results concerning the interaction effect between Pair Distance and Familiarity are similar to Experiment 5a.

5.8. General discussion

In the current study, I explored the relationship between familiarity and sensitivity to change using the well-controlled PCA transformations used in the previous experiments (Experiment 4a and b). There was a main effect of perceptibility of change (in number of components) between the response pairs when the change size was 4 SD. Additionally, there was a significant main effect of Familiarity and an interaction between Familiarity and perceptibility of change for the largest distance between pairs. This interaction means that the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces, but only on the largest distance between components. Surprisingly, this was not dependent on the magnitude of the change, as both Experiment 5a and 5b showed a similar pattern of results.

A result that stands out here is the difference in accuracy between familiar and unfamiliar faces for Pair Distance 2 in Experiment 5b, which was not present in Experiment 5a. Pair Distance 2 represent two pairs in the experiment: a pair of altered images in which one image had 3 components changed and the other 5 components changed, and the pairs in which one image had 5 components changed and the other 7 components changed. Taking a closer look at the data by looking at the accuracy for each of the pairs, shows us the difference is driven by a lower performance on unfamiliar faces for a pair with 3 and 5 components than familiar faces (accuracy is 0.61 and 0.73 respectively; Table 4 provides an overview of the accuracy per pair for familiar and unfamiliar faces). So, the question is, why might this be?

Table 4. averages per pair and Familiarity

Pair	Average accuracy unfamiliar	Average accuracy familiar	Pair distance
2 – 3	0.77	0.77	1
2 – 5	0.83	0.83	3
2 – 7	0.88	0.96	5
3 – 5	0.61	0.73	2
3 – 7	0.83	0.90	4
5 – 7	0.71	0.71	2

In Experiment 4a, an image with 3 components changed compared to the original had an average similarity rating of 4.06: 3.79 and 4.32 for unfamiliar and familiar faces respectively, and an image with 5 components changed compared to the original had an average similarity rating of 3.25: 3.02 and 3.48 for unfamiliar and familiar faces respectively. Thus, for unfamiliar

faces the comparison between images with 3 and 5 components changed, meant comparing images from close to the identity-boundary (with an average similarity rate of 3.79 and 3.02 respectively). On the other hand, for familiar faces the comparison between images with 3 and 5 components changed, meant comparing one image that one that was within the identity-boundary enough to guess that it was the identity to one that was close to the identity-boundary (with an average similarity rate of 4.32 and 3.48 respectively). Following the work of Beale and Keil (1995) that shows unfamiliar faces have a more gradual transition between identities around the identity-boundary than familiar faces, this finding makes sense, because the distinction for unfamiliar faces would be particularly difficult to make.

Table 5. average similarity rating per components changed

Components changed	Average similarity rating	Average similarity rating Familiar faces	Average similarity rating unfamiliar faces
2	5.11	5.29	4.93
3	4.06	4.32	3.79
5	3.25	3.48	3.02
7	3.11	3.30	2.91

However, this argument must be taken with caution, as a new procedure was used to create more natural-looking stimuli compared to the original stimuli used in Experiment 4a. Thus, the similarity ratings might be slightly different based on the new procedure. However, the change magnitude should still be more than sufficient to alter perceived identity for the transformations with higher number of components changed. Thus, it appears that the identity-boundary does influence the relationship between Familiarity and Pair Distance, but only for pairs close to the identity-boundary.

In sum, in the current experiments, there was a main effect of perceptibility of change (in number of components) and familiarity between the response pairs. Additionally, there was an interaction between Familiarity and perceptibility of change for the largest distance between pairs. Lastly, it appears that the identity-boundary does influence the relationship between Familiarity and Pair Distance, but only for pairs close to the identity-boundary. So far, the same components have been changed in order to control for identity changes, while change was defined by number of components (with the same base components) and magnitude of change (in SD). To further explore how PCA can be used to alter perceived identity, I will explore the effect

of different components in PCA space and how they relate to the identity-level representation: does changing different sets of components lead to different effects? In the following chapter, I will further explore sensitivity to change. Specifically, I will explore whether the visibility of a change of a particular component depends on the identity of the face, and if there are effects of familiarity.

6. Chapter 6 - Finding the category boundaries of faces: familiarity and the distinctiveness of transformations

Chapter main message

This chapter addresses a limitation of the previous studies, namely the use of the same set of principal components to alter faces. In this exploratory study, I will explore the effect of varying the principal components used in the transformations on perceived identity (distinctiveness of transformation) in order to better understand the transformations used in this thesis so far. More specifically, I propose a 4-AFC paradigm that compares the effect of different principal components and test whether it can be used to select more distinctive PCA components. The secondary aim of this chapter is to further explore the effect of familiarity on sensitivity to change. Based on the pattern of responses in Experiment 6b, there were no clear indicators that familiarity influenced the distinctiveness of transformation.

6.1. Introduction

In the previous chapter, an interaction effect between sensitivity to change and familiarity was found, if the change is big enough. Surprisingly, it was also found that magnitude of change does not appear to affect the role of familiarity in sequential face matching. However, due to the clear limitation that the same set of components in PCA space are changed in the studies so far, it is unclear whether this effect generalizes to transformations using other principal components. In the previous studies 6 and 8 identities are used, and there is a possibility that some components did not have a very noticeable effect on one identity but do have an effect on another, i.e., there could be a distinctiveness of transformation. In this chapter, I further explore the role of distinctiveness of transformation and assess whether it can be used as a paradigm to study effects of familiarity.

One reason why different principal components can induce a different change in one identity versus another is the idiosyncratic nature of faces. In the review of Chapter 1, a brief overview of how faces vary was given. Variability in images of faces can be divided into within- and between-face variability. Within-face variability is all the variability that is present in all different instances of the face in 1 identity, and between-face variability is the variability that defines differences between different identities. Both within-face and between-face variability are idiosyncratic, i.e. distinctive for the individual, and subsequently, both affect identification.

Both types of face variability can be captured with PCA, dependent on the set of images that are given as input. In the studies presented in this thesis so far, within-face variability is controlled for and the focus is on between-face variability. I want to note that it is beyond the scope of this thesis to include within-face variability and I therefore work with well-controlled, frontal images of averages, but acknowledge the importance of understanding the role of it in learning faces and real-life face recognition (for an overview of reasons why, and fundamental research in this field, please see (Bindemann & Hole, 2020; Mike Burton, 2013)). To acknowledge the importance of within-face variability, the designs used in this thesis can be adapted to explore within-face variability in future research (for example, in a similar fashion as in (Burton et al., 2011)). Although the transformations are well-controlled and aim to influence identities in a similar fashion, the information stored in a participant's internal representation on within-face variability could still influence the effect a particular transformation might have.

To change perceived identity, one presumably must move the face outside the area in face space in which the individual is represented. However, little research has been done on the boundaries of an individual's face space. One of the few studies on this topic has been done by (Beale & Keil, 1995), which I briefly discussed in Chapter 0. To summarize, the authors explored the effect of familiarity on the categorization of faces by using morph sequences between two different individuals' faces and found that familiar faces have a sharper decision boundary than unfamiliar faces. Although this study provides some insight in the boundary of the area in which an individual's face is represented in face space, one obvious limitation is that it only provides information about the area of face space that lies between the two identities the morph is based on. In other words, the boundary is only explored in comparison to another identity, forcing the boundary to end at the overlap with that other identity. However, as discussed above, within-face variability can share some overlap with other identities and the boundary is decided by experience with that variability. Thus, the boundary between two identities is not the same as the boundary of an individual's region in face space. For example, in the (Beale & Keil, 1995), one morph was from Kennedy to Clinton. One could learn then what components are relevant for Kennedy, but only relative to how it distinguishes him from Clinton. Although insightful, it remains unclear whether or not these components are the same or most informative as to what makes Kennedy's face different compared to other faces in face space. Additionally, distinctiveness influences how morphs are perceived (Angeli et al., 2008): the decision boundary between two morphs of a typical to distinctive face shifts so that the mid-point will 'look like' the distinctive face.

Beale and Keil (1995) matched identities on perceived familiarity, distinctiveness, apparent age and image quality, further suggesting an effect of these factors on the level of morph needed to perceive the transition into another identity. The effect of distinctiveness on face recognition has been reasonably well established (e.g. Angeli et al., 2008; V. Bruce et al., 1993; Hancock et al., 1995; Valentine, 1991; Wickham et al., 2000), though there is no formal definition of distinctiveness. There are two common operational definitions used in research on distinctiveness: defining a distinctive face as standing out in a crowd, and the extent to which a distinctive face is different from a typical/average face. Whereas the second definition provides a clearly measurable definition of distinctiveness, the first is less explicit in what distinctiveness means. Wickham, Morris and Fritz (2000) describe that a face can be distinctive in terms of

standing out in a crowd for multiple reasons: “for example, a single or a few distinctive features, a similarity to a familiar face, or an unusual expression.” Thus, although it is highly likely that distinctiveness has an effect, it remains unclear what makes a face distinctive in the first place.

In the current study, I aim to gain a better understanding of the effect of different components on perceived identity – i.e. distinctiveness of transformation – in order to better understand the transformations used in this thesis so far. For example, a transformation that affects lips might vary in effect depending on the individual's lips, e.g. changing the lips of Angelina Jolie will likely have a different effect than somebody with less distinctive lips to begin with. Specifically, I explore if PCA can be a possible technique to find distinctive components that are based on a wider range of variability (as our set is made up of 72 identities total) opposed to the previously mentioned morphing techniques. In this study, a distinctive component is a component that changes the face to the greatest extent compared to other transformations. Since a distinctive component is defined as a component that changes the face to the greatest extent compared to other transformations in this study, the question to the participant will be focussed on identifying the largest change in an appearance. Thus, the current study will ask the question “Which of the four changed pictures you think has CHANGED the appearance of the person the MOST?”. Since the emphasis of this study is finding out the effect of different single components, the study will be purely exploratory and will refrain from formal statistical analysis (as capturing the effect of distinctiveness in a way that is generalizable would require more identities and more transformations, and mostly likely a very large sample).

A secondary aim of this study is to explore any differences in distinctiveness of transformation between familiar and unfamiliar faces. As discussed to a great extent in the thesis so far, it remains unclear what the differences in internal representations of a familiar and unfamiliar face are. Due to this unclarity, it is difficult to predict how familiarity will influence or interact with distinctiveness of transformations. Purely speculating, there are 3 scenarios: familiar faces are more consistent than unfamiliar faces in which component is most distinctive, the reverse could be true, or there is no difference evident in the current paradigm. I will briefly provide arguments for each scenario.

As discussed in Chapter 2, in every type of task in which visual information is encoded and/or compared, either to an internal representation or to an externally displayed source, visual working memory (VWM) is involved. VWM is a component of working memory. VWM has a

limited capacity (Luck & Vogel, 1997; Ma et al., 2014), meaning there is only a finite amount of visual information at a certain moment in time that we can encode when perceiving a face. Additionally, there is a theory that the default of the human brain is to minimize the costs involved in any process (e.g. the Free energy principle, Friston, 2009, 2010; Sengupta, Stemmler, & Friston, 2013), making it unlikely that the visual system will keep collecting visual information on features indefinitely. From VWM, the information moves for further processing. There is a bidirectional transfer of information between working memory and long-term memory (for a review, see Burgess & Hitch, 2005).

In the first scenario, familiar faces are more consistent than unfamiliar faces in which component is most distinctive, the reasoning is similar to when familiarity could aid face perception. Familiarity could aid the comparison of two faces by using the memory of a familiar face to enrich the image currently in visual working memory (VWM) with more details to compare. In the context of the current paradigm the enriched internal representation of a familiar face would mean a greater probability of having a visual representation of distinctive features. This would subsequently make a change to a face more evident compared to an unfamiliar face that does not have an internal representation. Thus, a more familiar face with a better representation of that face in memory will contain more information that can be loaded into VWM and that should allow for a better comparison to what is presented externally.

In the second scenario, unfamiliar faces are more consistent than familiar faces in which component is most distinctive, the reasoning is similar to when familiarity disrupts face perception: this ‘enrichment’ for familiar faces has the potential to disrupt the comparison, by overwriting VWM content of the presented image with the visual information retrieved from the memorized face. In this situation, the participants familiar with the face would be less sensitive to change, because it is minimized in their VWM by information from LTM. However, this assumes that a presence of a stored internal representation, i.e. familiarity, relates to distinctiveness. If the latter is not true, both familiar and unfamiliar faces will be equally consistent in their effects of the components. The unclear predictions of the effects of familiarity on distinctiveness of transformation are another reason for the exploratory nature of this study and its statistical analysis. In other words, the goal is not to provide a conclusive answer what the relationship is between familiarity and distinctiveness of transformation, but whether or not this paradigm can be used to explore this question.

The results of Experiment 4a (Somai & Hancock, 2022) showed that the effect on perceived identity of single- and multidimensional changes was very similar, on average. However, only a small range of components were selected and either single component changes were made or multidimensional changes with the same components changed (components changed: 10 – 11, 10 – 12, 10 – 14, and 10 – 16). It remains unclear if the number of components always effect perceived identity the same, or if the perceived change is dependent on the other components transformed. E.g., perhaps, one component darkens eyebrows another changes their shape, in concert they may have a bigger effect than when paired with a different component.

In the current chapter, I will explore the effect of varying the principal components used in the transformations on perceived identity in order to better understand the transformations used in this thesis so far. Additionally, the secondary aim of this chapter is to explore the effect of familiarity on sensitivity to change as defined by consistency in choosing a transformation in a set of multiple transformations that are all made up of the same number of components and magnitude of change.

6.2. Methods Experiment 6a

Participants

We tested 60 participants (41 female, 19 male, 0 other), mean age = 29.4 years, 1 missing values, SD = 9.80 age range = 18 - 58 years. Testing was done through an online platform, Testable, recruitment was done through Prolific. Participants were naive to the purpose of the experiment. All participants acknowledged informed consent after a written explanation of the procedure. Participation received monetary compensation. The experiment was approved by the local General University Ethics Panel of the University of Stirling.

Apparatus and stimuli

The experiment was programmed through the online platform Testable and participation was only allowed via a PC or laptop. A calibration to control stimulus size was performed in form of matching the size of an on-screen bar to a credit card, to ensure proper visualisation throughout the experiment. The stimuli creation followed a similar procedure to that used in Experiment 4b (Chapter 4), with the important difference that other components were varied. Each identity had

a total of 8 transformations, each of 4 different components changed (referred to as the variable components) in combination with both 1 and 4 other principal components (referred to as the base components; for an example, please see Figure 35). The base component is component 10 in the 2-component condition, and components 10, 15, 16, and 17 in the 5-component condition. So, for example, if component 11 is the variable component, this would mean a transformation of components 10 and 11 in the 2-components condition and 10, 11, 15, 16, and 17 in the 5-component condition.

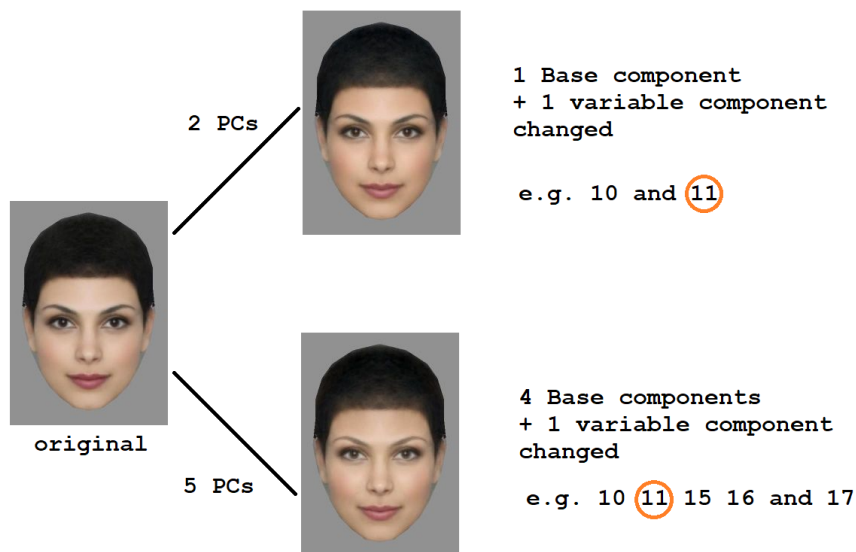


Figure 35 Example of the transformations used.

In the 2-component condition, components 10 is always changed (base component), and in the 5-component condition 10, 15 16 and 17 are changed, in addition to the variable components (circled in orange). In this study, the effect of component 11, 12, 13 and 14 were explored.

Procedure

The experiment consisted of a familiarity check and a face matching task. The same familiarity check as previous experiments was used: participants were asked “Would you recognise this person on the street? yes(y)/ no(n)” for each identity (presented in a random order). Additionally, a follow-up question was added: “How sure are you that you do/do not recognize this person?” Participants had the following response options: 1- I do not know them; 2- they look familiar but I cannot place them; 3- I recognise them but cannot remember details (such as name); 4- I know who they are; 5- I know who they are and I am very familiar with their content. Afterwards,

participants were presented with the face matching task instructions and had 2 initial practice trials with an identity not included in the actual experiment. This was followed by 32 trials in a random order. In each trial, the original was shown on the top of the screen with 4 transformed images below (see Figure 36 for an overview of a trial). The participants were presented with the question “Which of the four changed pictures you think has CHANGED the appearance of the person the MOST” with four pictures as response options that could be selected by simply pressing the corresponding number on the keyboard (i.e. 1 for the first picture). Response was possible after 5 seconds. Each participant saw all these variations divided in trials of 2 and 5 PCs changed. The original images was depicted at the top and the four variations: any one of the base components – 11, 12, 13, 14 – in combination with 1 variable component – 10 – in the 2 PCs conditions, or 4 variable components – 10, 15, 16, 17 – in the 5PCs condition (for an overview of the trial and an example of the stimuli for both 2PCs and 5PCs changed, see Figure 36 and Figure 37). The experiment showed identities in a random order and the order the transformations were presented was counterbalanced within identity.

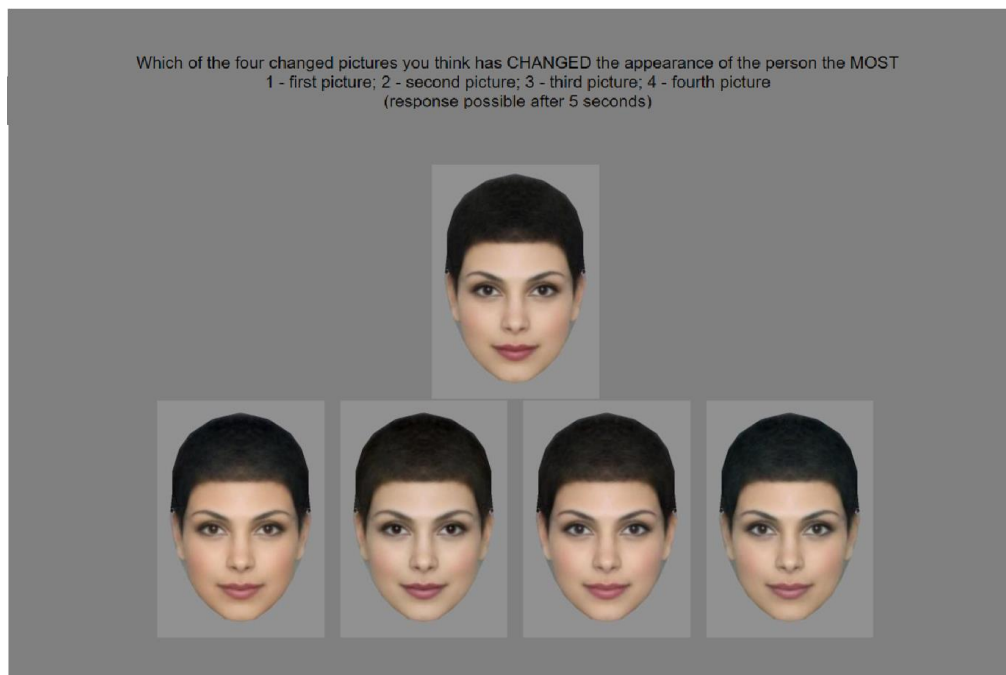


Figure 36 Example of a trial in which 2 components are changed (2PCs condition).

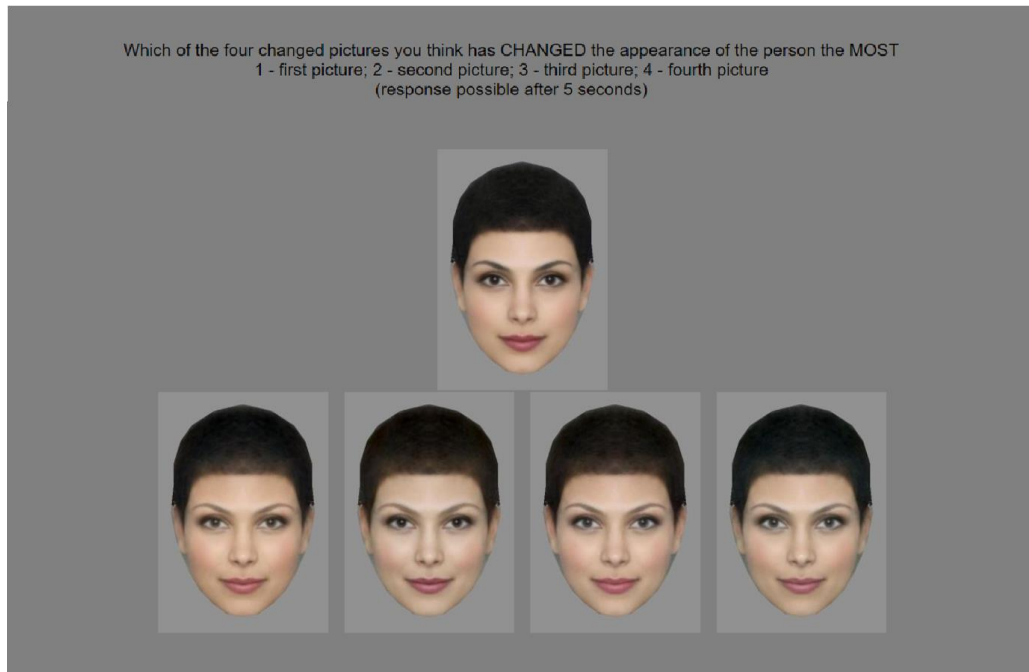


Figure 37 Example of a trial in which 5 components are changed (5PCs condition).

Analysis

No formal statistical analysis will be performed due to the exploratory nature of this experiment. Instead, the focus will lie on descriptive statistics such as frequency and proportion.

6.3. Results Experiment 6a

The frequency of components selection is given per ID, change size and Familiarity in Figure 38. However, due to ID 1 and 3 being relatively well-known compared to ID 2 and 4 who are relatively unknown, these graphs can be difficult to compare. To enable an easier comparison, the same data described in terms of proportion of the components is depicted in Figure 39.

For ID1 it is clear that the component 12 was always the most chosen one, regardless of familiarity. ID 2 shows a stronger effect of component 14 when the change is large but no clear effect of familiarity. In ID3 the component 14 appears to be chosen more often for familiar faces, and little preference in the smaller change when unfamiliar, and a preference of component 11 in the large change when unfamiliar. The pattern of ID4 is similar to ID2, showing a stronger effect of a component when the change is large but no clear effect of familiarity. However, in ID2 the most chosen component in the larger change was component 14 and in ID4 it is component 12.

Additionally, in ID4 component 11 is chosen more often in the small change condition when unfamiliar with the face, but no such effect is seen when familiar with the face.

Looking at the frequency plots (Figure 39), it is clear that more participants were familiar with ID1 and ID3 than unfamiliar. The opposite is observed for ID2 and ID4, which are unfamiliar to more participants than familiar. The levels of familiarity appear similar for all identities (see Table 6), overall, the level of familiarity for familiar faces was 3.60 and for unfamiliar faces 1.48. The value 3.6 is between ‘I recognise them but cannot remember details (such as name)’ and ‘I know who they are’, and 1.48 is between ‘I do not know them’ and ‘they look familiar but I cannot place them’.

Table 6 Familiarity rating per ID

ID	Familiar	Unfamiliar
1	3.69	1.56
2	3.25	1.25
3	3.94	1.9
4	3.53	1.22

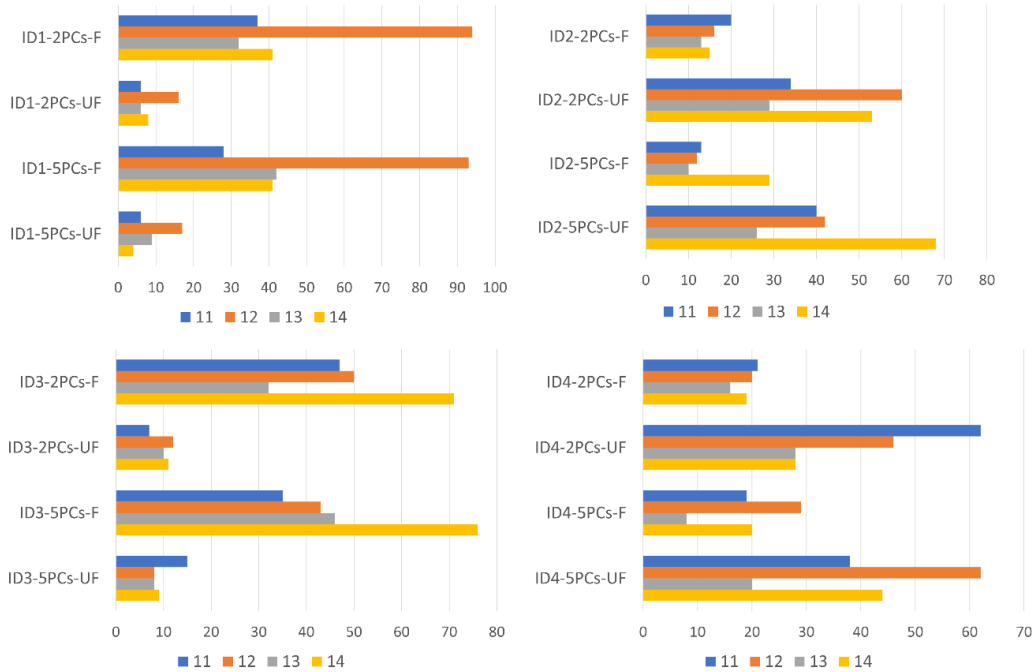


Figure 38 Components selection (frequency) per ID, change size and Familiarity.

Per ID the top rows are the 2 component changes and the bottom are 5 components changed. 'F' is to indicate familiar faces and 'UF' indicates unfamiliar faces.

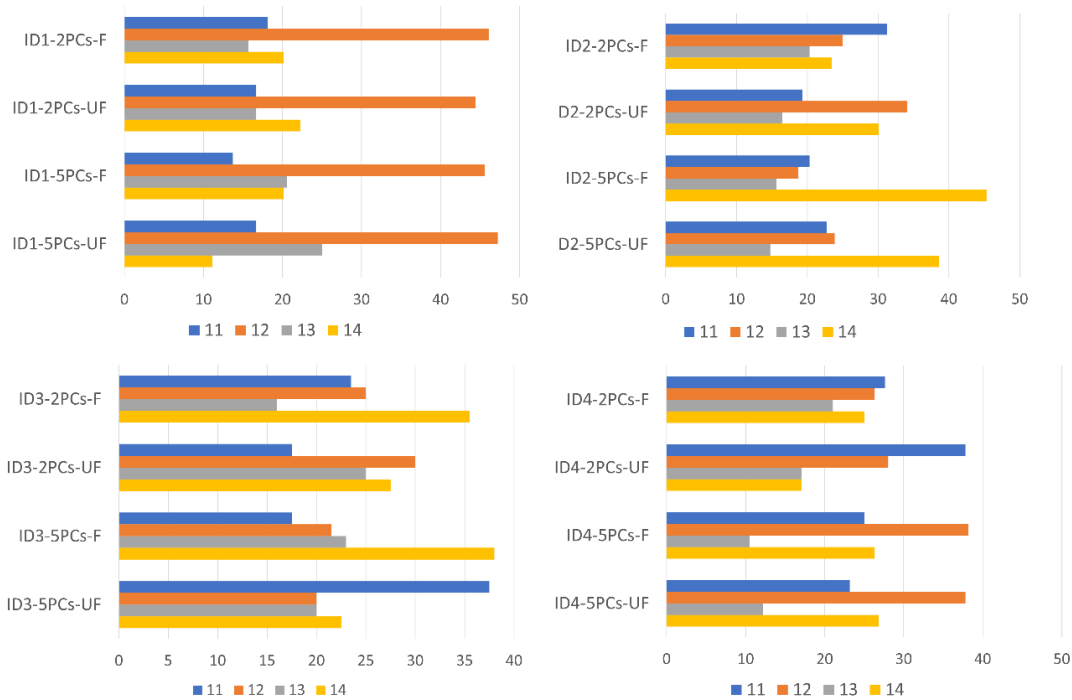


Figure 39 Components selection (%) per ID, change size and Familiarity.

Per ID the top rows are the 2 component changes and the bottom are 5 components changed. 'F' is to indicate familiar faces and 'UF' indicates unfamiliar faces.

Post-hoc analysis of the Gabor distances and DNN similarity scores

As discussed in Chapter 0, there is a (linear) relationship between perceived similarity and Gabor space and DNN similarity scores. To check if the preference for certain components can be explained away by the size of the change in an image-based space, the Gabor distances and DNN similarity scores are calculated for all the transformations. Neither the Gabor distances nor DNN similarity seem to explain the found results in an obvious way, i.e., the components with the largest Gabor distance does not always equal the preferred component in the behavioural data.

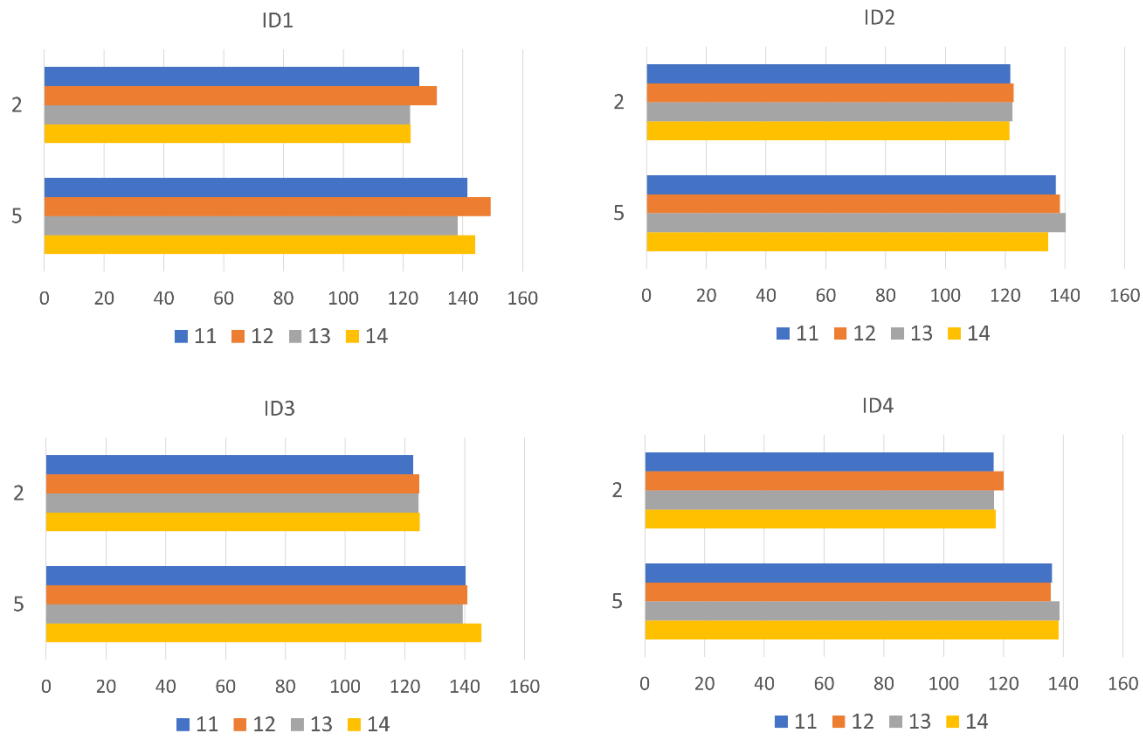


Figure 40 Gabor distances per ID, change size and component.

Per ID the top rows are the 2-component changes and the bottom are 5 components changed.

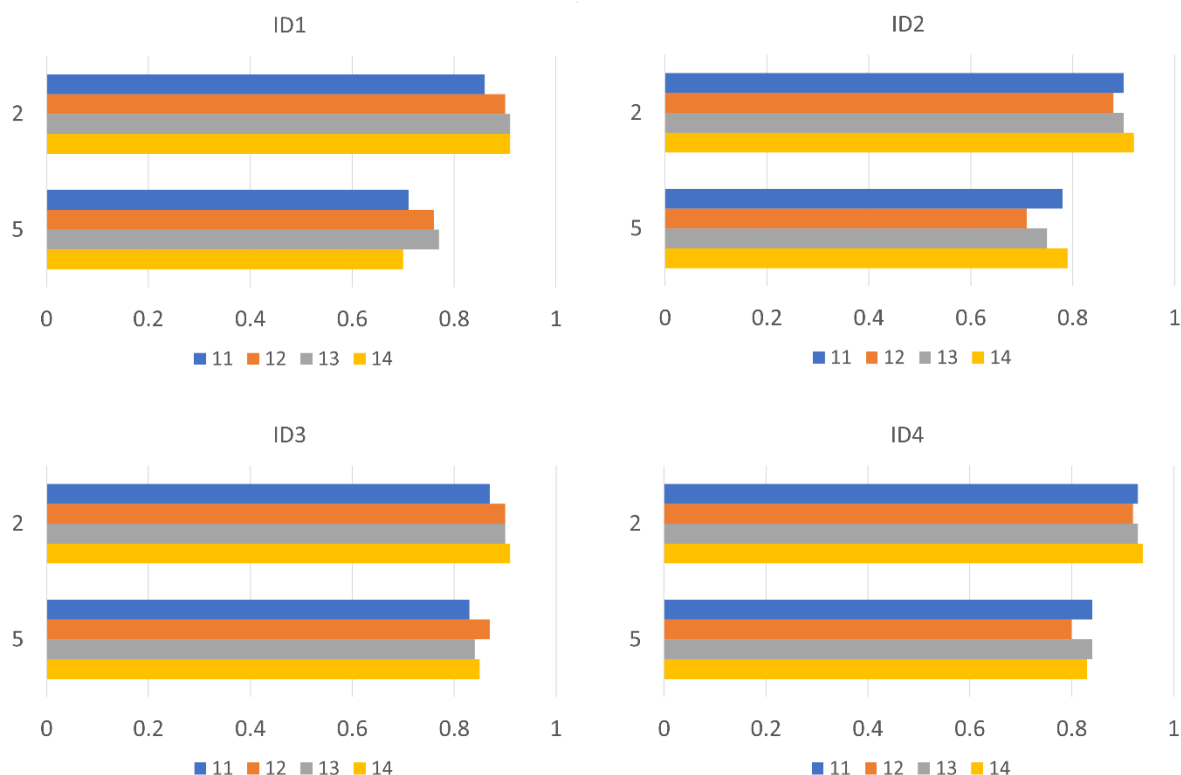


Figure 41 DNN similarity scores per ID, change size and component.

Per ID the top rows are the 2-component changes and the bottom are 5 components changed.

6.4. Discussion Experiment 6a

In this first experiment, I explored the effect of varying the principal components used in the transformations on perceived identity in a 4-AFC paradigm. Additionally, the secondary aim of this experiment was to test the suitability of the proposed paradigm for exploring the relationship between familiarity and distinctiveness of transformation. As expected, it is difficult to provide an indication of what the relationship is between familiarity and distinctiveness of transformation. However, there appears to be a preference for a certain component dependent on the identity, one that cannot be reasoned away by the size of the change measured in an image-based space (Gabor distance). Thus, these findings do indicate this paradigm picks up on idiosyncratic variability, and it seems a promising first step towards exploring distinctiveness of transformation, and its relationship to familiarity.

The effect shown in ID3 when looking at the proportion of PCs selected is what fits best with the current literature on the effect of familiarity: when one is familiar with the face, the

perceiver is more sensitive to change in a particular component, in this case component 14. However, this pattern was not found in the other identities. One of the biggest limitations in this study was the uneven distribution of participants that were familiar and unfamiliar with certain identities. Thus, before concluding anything about the relationship between distinctiveness of transformation and familiarity, it is important to address this limitation first.

Another noteworthy observation in Experiment 6a is the comparison between ID1 and ID3. As mentioned above, ID3 had a clear component having the largest effect when perceivers considered it a familiar face, but not when one was unfamiliar with the face. In contrast, ID1 had a clear component that, regardless of familiarity, had the strongest effect on perceived similarity. The post-hoc exploration of the Gabor distances and DNN similarity scores did not provide an image-based reason for this.

In sum, it appears the PCA transformations could be used to explore familiarity and the distinctiveness of change. However, there appears to be an effect of identity, thus the sample size and/or design should be aimed to account for the imbalance between participants who are (un-) familiar with the identity. These limitations were addressed in Experiment 6b.

6.5. Methods Experiment 6b

To address the limitation of the uneven familiarity distribution over our identities (i.e., the number of people familiar or unfamiliar with the identities), I chose 2 Spanish and 2 Canadian celebrities and targeted 60 participants from the respective countries.

Participants

I tested 120 participants (55 female, 64 male, 1 other), mean age = 33.5 years, SD = 10.9 age range = 18 - 68 years. Testing was done through an online platform, Testable, recruitment was done through Prolific. Participants were naive to the purpose of the experiment. All participants acknowledged informed consent after a written explanation of the procedure. Participation received monetary compensation. The experiment was approved by the local General University Ethics Panel of the University of Stirling.

Stimuli

The creation of the stimuli followed a similar procedure as those used in Experiment 6a, with the important difference that other identities were used in the calculation of the PCA space. Changing the identities in the PCA space was necessary to make the changes on the new identities look natural. The four new identities (the 2 Spanish and 2 Canadian celebrities) were added and four others, chosen at random, were removed. Thus, the specific components that were varied will not reflect the same dimensions, as this is dependent on the PCA space used.

Procedure

The experiment followed a similar procedure to Experiment 6a, barring the identities presented.

6.6. Results Experiment 6b

The frequency of components selection is given per ID, change size and Familiarity in Figure 42. Again, to enable an easier comparison between identities, the same data described in terms of proportion of the components is depicted in Figure 43. Unlike Experiment 6a, all identities appear to have the same component with the strongest effect on appearance: component 12. Note that component 12 in Experiment 6a and 6b are not the same, as these are dependent on the PCA space. The only exception to the preference for component 12 is the 5PC transformation in unfamiliar faces for ID2 and ID4, and the 2PC transformation in familiar faces in ID2 and ID3. The levels of familiarity appear similar for all identities, except for ID1 which has a lower level of familiarity (see Table 7), overall, the level of familiarity for familiar faces was 3.20 and for unfamiliar faces 1.35.

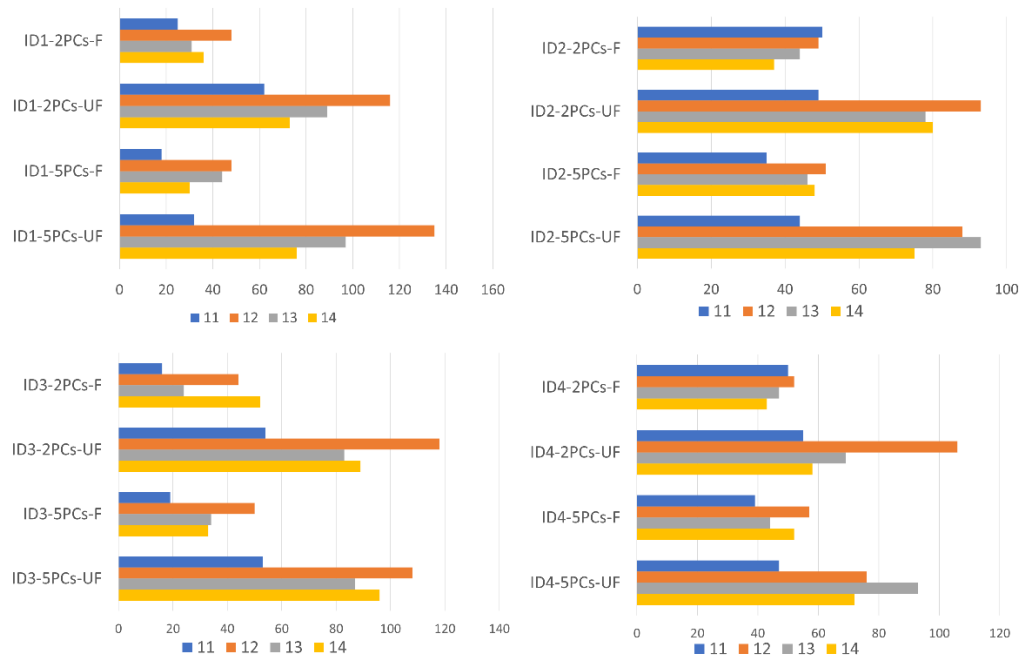


Figure 42 Components selection (frequency) per ID, change size and familiarity.

Per ID the top rows are the 2 component changes and the bottom are 5 components changed. 'F' is to indicate familiar faces and 'UF' indicates unfamiliar faces.

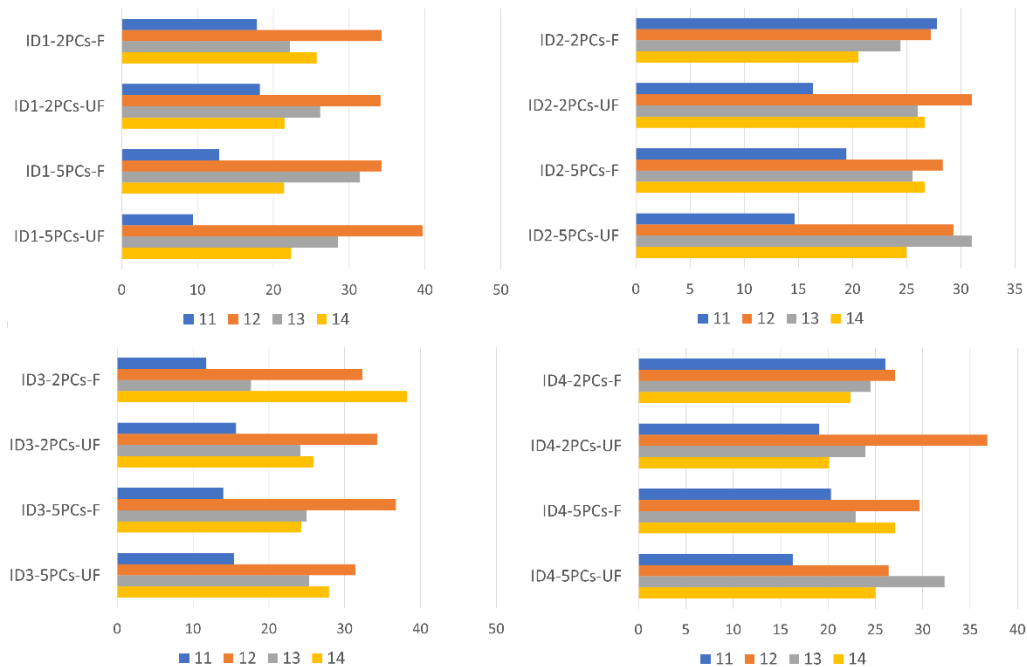


Figure 43 Components selection (%) per ID, change size and familiarity.

Per ID the top rows are the 2 component changes and the bottom are 5 components changed. 'F' is to indicate familiar faces and 'UF' indicates unfamiliar faces.

Table 7 Familiarity rating per ID

ID	Familiar	Unfamiliar
1	2.69	1.46
2	3.58	1.31
3	3.53	1.29
4	3.02	1.35

Post-hoc analysis of the Gabor distances

Similar to Experiment 6a, the Gabor distances of the transformations are checked to see if the preference for certain components can be explained away by the size of the change in an image-based space. Again, the preference for component 12 in the behavioural data does not seem to be because of a greater image-based difference, as the Gabor distance for component 12 transformation is similar to the other components.

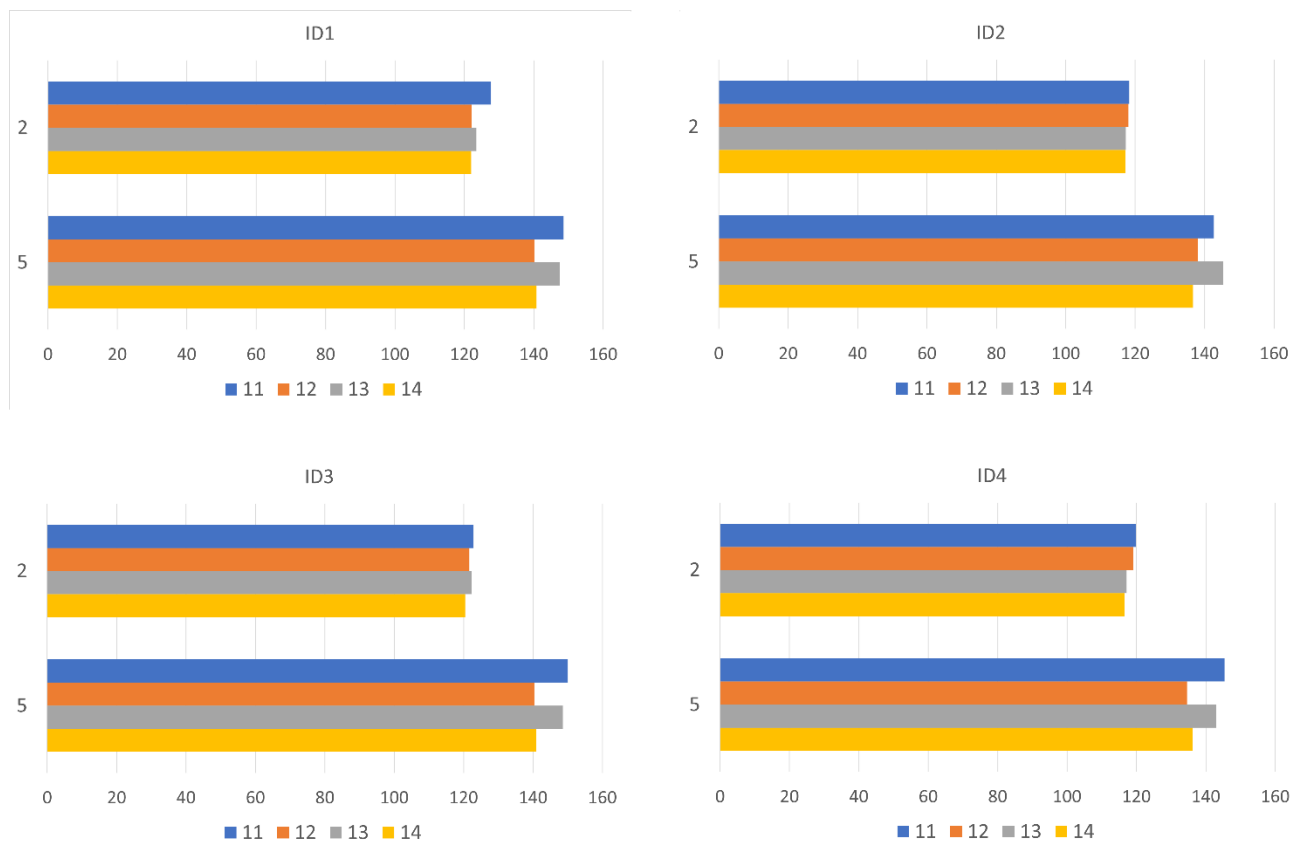


Figure 44 Gabor distances per ID, change size and component.

Per ID the top rows are the 5-component changes and the bottom are 2 components changed.

6.7. General discussion

In this study, I explored the effect of varying the principal components used in the transformations on perceived identity in order to better understand the transformations used in this thesis so far. More specifically, I proposed a 4-AFC paradigm that compares the effect of different principal components to explore if the effect of components is different per identity and if I can use a 4-AFC paradigm to select more distinctive PCA components. The secondary aim of this chapter is to test this proposed paradigm to explore the effect of familiarity on sensitivity to change. In the current paradigm, I defined sensitivity to change in a novel way, by defining it as consistency in choosing a transformation in a set of multiple transformations that are all made up of the same number of components and magnitude of change.

The results of Experiment 6a could not be interpreted due to the distribution of familiarity: 2 identities were familiar for the majority of the participants and the two others were mostly unfamiliar. This makes it impossible to differentiate between the effect of familiarity and any idiosyncratic effects. For this reason, the experiment was repeated with other identities in Experiment 6b. Although there appeared to be a preference for a certain component dependent on the identity in Experiment 6a, this was not found in Experiment 6b. The results of Experiment 6b show that component 12 has the strongest effect on appearance in all identities. Since the same component affected all identities in a similar fashion, there were no clear indicators that familiarity influenced the distinctiveness of transformation.

The finding that component 12 has the strongest effect on appearance in all identities in Experiment 6b can be considered in line with the idea of Wickham et al. (2000) that a face that can be mostly typical and only has a few distinctive components can still be perceived as distinctive. In other words, component 12 might be a facial feature or a set of features that Wickham et al. (2000) refer to as ‘psychologically distinctive’ but not necessarily opposite of typical face, where typical is defined as average. Due to this diverse nature of distinctiveness, it is difficult to interpret the findings in this study in a clearcut fashion.

As mentioned above, the main aim of this study was to better understand the transformations used in the thesis so far by exploring if the effect of components was different per identity and different in perceived change. Based on the findings in Experiment 6b, it appears that none of the principal components tested in this study had an identity-specific effect.

However, due to several limitations of the study, it is difficult to hypothesize whether this finding will generalize beyond this particular design.

Firstly, the research question used in this study is focussed on dissimilarity opposed to similarity, since the questions presented to the participants was the following: “Which of the four changed pictures you think has CHANGED the appearance of the person the MOST?” Since the findings of both studies are not in line with previous work in this thesis, in particular the lack of relationship between the image-based Gabor distance and perceived similarity discussed in Chapter 0, one could deduce that the relationship between perceived similarity and image-based spaces does not imply the opposite relationship between perceived dissimilarity and image-based spaces (Hancock et al., 2020). However, without further research, it is unclear to what extent the different nature of the question and research design have affected the findings.

Secondly, an obvious practical limitation of this paradigm is that the transformations are specific to the PCA space calculated, and that in this study different PCA spaces per experiment were used. This means it is not possible to directly compare the transformations used in Experiment 6a to the transformations used in 6b, which can make it difficult to generalize these findings (to other identities or transformations). If this were the aim, one would need a large PCA space based on a large number of identities, so that any new identities can be well represented within the existing PCA space. However, it remains unclear how many identities are necessary to make this possible. Additionally, the work involved in creating a PCA space suitable for these experiments (10 frontal images of good quality and relatively high resolution go into 1 average), would be extremely laborious.

To overcome the limitation of generalizing to other identities and ensuring all identities can be represented as well as possible, Deep Neural Networks (DNNs) can be used. DNNs capture all the information in thousands to millions of images and creates a representational space based on all the image features it can extract from the images that are given as an input. In Chapter 7 - Review on the use of deep neural networks in understanding human face recognition’, I will provide an overview of why and how DNNs are used to better understand human face recognition. Additionally, developing a DNN that shows human-like familiarity could potentially be used to select the most defining features of a face. The idea of human-like familiarization of faces in a DNN is explored in three computational experiments (Chapter 8 - Exploring the

representation of familiar faces using deep neural networks.), in order to better interpret these findings in the context of how familiarity affects our internal representation.

In sum, this study explored a 4-AFC paradigm that compares the effect of different principal components in order to better understand the transformations used in this thesis so far. The secondary aim of this chapter was to test this proposed paradigm to explore the effect of familiarity on sensitivity to change, defined as consistency in choosing a specific transformation over other transformations that are all made up of the same number of components and magnitude of change. The results of Experiment 6a could not be interpreted due to the distribution of familiarity and the results of Experiment 6b show that component 12 has the strongest effect on appearance in all identities and no apparent effect of familiarity on the distinctiveness of transformation.

Part II

7. Chapter 7 - Review on the use of deep neural networks in understanding human face recognition

Chapter main message

As an introduction to the second – computational – part to this thesis, I discuss the most relevant contributing factors in the development of the field of computational cognition, and why in particular deep neural nets are suitable to explore human face recognition. I end this chapter by reviewing the handful of studies that applied deep neural nets to explore the internal representation of faces.

7.1. Introduction

So far, I have only discussed the psychological approach in tackling the question of effects of familiarity in human face recognition. However, the different perspectives on the question “Are faces special?” discussed in Chapter 1 enunciate the true issue in face research, which is the complexity and mystery that is still involved in face processing. Humans are extremely skilled in face perception and recognition, and ceiling effects make it challenging to deduce any information of the complex underlying mechanisms based on behavioural, neuroscientific and psychophysical data.

The problems of complexity, untangling interactions and deconstruction is a problem faced in many disciplines of psychology. In the 1950s a new subfield developed that tackled a selection of these problems using mathematical models: Mathematical Psychology (For a review, see Batchelder, 2010). Mathematical psychologists realized that mathematical notions stemming from information sciences can be used for cognitive modelling, allowing us to gain a deeper understanding of data in psychology. Recent technological advances are likely to be related to the current renewed uprise of mathematical psychology that led to the growing field of computational cognition. The abundance of studies on face perception has given us a wealth of data and testing paradigms that can function as the basis of the next step: an integrative approach in the understanding of face perception that aims to unify previously collected data.

In this chapter, background will be provided on how recent technological and theoretical advances lead to the evolution of tools and paradigms that is now collectively known as computational cognition. Additionally, following a chronological timeline, I will discuss

experiments which applied deep neural networks (DNNs) that have contributed to our fundamental knowledge on face processing. In this review, I limit the discussion on computational experiments to those in which deep neural networks have been used to gain a greater understanding in human face recognition. For a wider overview of computational models beyond the topic of familiarity, please see O'Toole & Castillo (2021). To avoid confusion, I will explicitly refer to human and artificial face recognition, in which artificial face recognition could refer to any of the stages mentioned in the following subsection.

7.2. Stages of artificial face recognition

As mentioned in Chapter 1, when discussing human face recognition, the term often reflects several different subprocesses (Ramon et al., 2019) and can refer to face perception, learning, identification and expression recognition. Artificial face recognition has a similar stage-wise division, and it has been described to have three stages: face detection, representation, and identification (Zhao et al., 2003). Although not an exact one-to-one match between human and artificial face recognition, there is an obvious parallel to be noticed between perception and detection, learning and representation, and identification in both. For this PhD project, I am interested in only the stage of representation. However, from the Bruce and Young model on human face recognition (Bruce & Young, 1986) it is clear that these stages influence each other, thus, a short description is appropriate.

For humans, detecting a face in a visual scene is an easy, low-effort phenomenon and faces seem to be a special category in human visual detection (for an overview of evidence, see e.g. 't Hart et al., 2011). Although automatic face detection has been a topic of interest for more than five decades in computer vision, the models struggled with image variation, such as differences in lighting, pose and distance (Zafeiriou et al., 2015). The lack of a functional automatic face detection meant that research with face images often included the laborious task of marking faces manually, resulting in studies with relatively few, well-controlled, lab images.

Major progress in this field was made with the seminal work of Viola and Jones (2001), with their application of boosted cascades of simple features. Briefly explained, a cascade is a method to combine complex classifiers, which allows for the focussing of attention on regions of an image important for classification, thus increasing processing speed. Viola and Jones

combined this method with a novel way of representing an image that needed very few computations per pixel, and a selection of essential features. Although there has been some progress made in the development of face detection algorithms based on machine learning, the most widely used face detection in applications are still based on the haar-cascades described in the work of Viola and Jones (2001).

The second stage of artificial face recognition is the creation of the face representation (Zhao et al., 2003). Similar to the progress in face detection, the progress in representation was also based on the principle of cascading. In 2012, performance of object classifiers greatly increased when ‘AlexNet’ showed the potential of cascading multiple layers of processing units that each dealt with a different level of extraction and transformation (Krizhevsky et al., 2012). I will discuss this work in greater detail in the subsection ‘Why now?’. This type of deep learning representation also bears resemblance to the human visual system, and increased the performance of artificial face recognition systems to near human-like levels (Wang & Deng, 2021). I will further discuss the content of the representation in subsection ‘Why DNNs’.

Face identification is the last stage of artificial face recognition. In this stage, the representational vector is compared to a database of stored representational vectors to match the face. The face is identified by finding the smallest distance between the input vector compared to all other vectors in the database. Alternatively, there is face verification, in which a one-to-one comparison is made and a threshold is used to verify the face identity. The most commonly used distance measure is the Euclidean distance and the cosine angle between two vectors. However, different measures exist, such as the Manhattan, Mahalanobis and Jaccard distance. For the experiments discussed in Chapter 8, Euclidean distance was used. The other measures will be discussed further in Chapter 9.

7.3 Why DNNs?

The reason why I – and many other researchers in the field of computational cognition and computer vision – focus on DNNs is two-fold: the characteristics of the model and the technological advances from the last decades. In this section, I will discuss the characteristics that make DNNs to be a particularly good candidate in the exploration of human face recognition, followed by a brief discussion of major developments and events in deep learning.

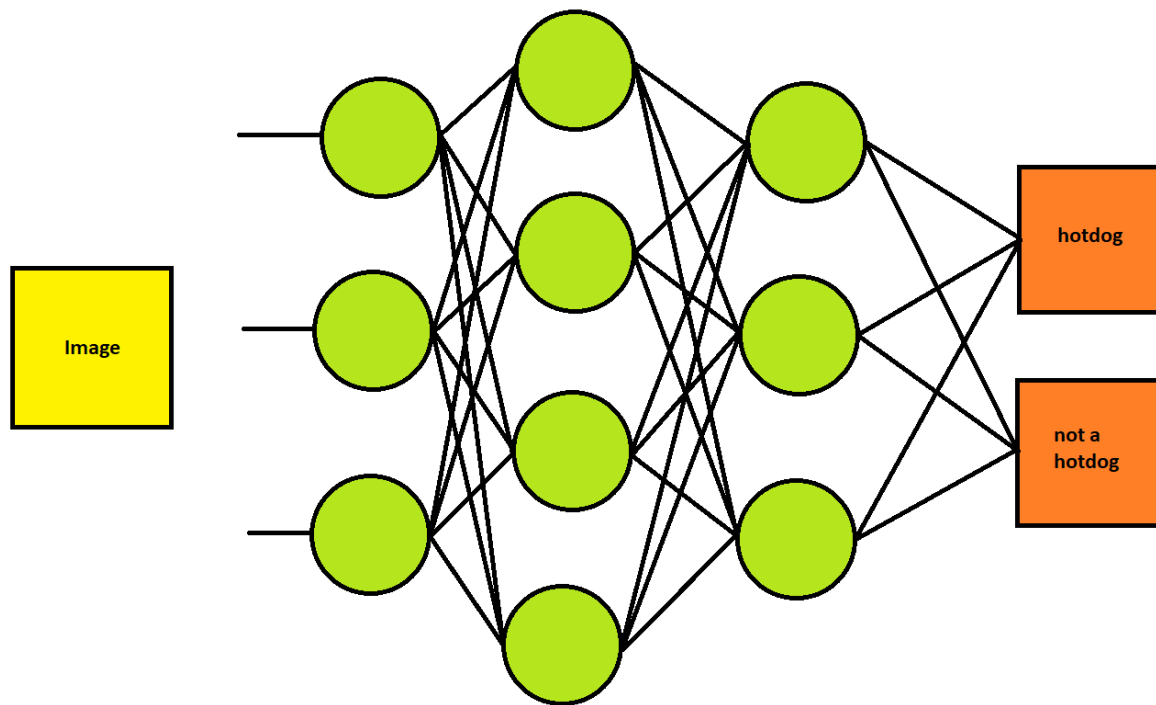


Figure 45. Schematic example of a DNN.

Before discussing reasons for using DNNs in research, however, I will provide a brief overview of what DNNs are.

As stated before, DNNs capture all the information in thousands to millions of images and creates a representational space based on all the image features it can extract from the images that are given as an input. In an extremely simplified description of events, a DNN take an image as input in the shape of a vector/matrix of values, which then undergo a sequence of computations by a variety of mathematical operations, and change the values to accommodate a task (such as classification). During training, the goal of a DNN is to minimize the difference between the predicted output and the actual output. Figure 45 shows an example of a very basic neural net. Each of the vertical rows of nodes are considered a layer. As input, each layer receives the output of the previous layer. The last layer in a DNN is often the ‘Softmax’ classification layer. This layer normalizes a vector of values using logistic regression, changing the values so that they follow a probability distribution of which the total of the values sums up to 1.

DNNs are biologically inspired systems. It should come as no surprise that machine learning, an important aspect of artificial intelligence, is inspired by the human brain because

humans have an extraordinary ability to learn. In this section, I will discuss several examples to illustrate this.

Frank Rosenblatt, who is considered one of the founding fathers of the field of deep learning (Tappert, 2019), was a psychologist who aimed to understand cognitive functioning through computational mechanisms. Specifically, he aimed to link the external physiological world to behaviour, in his own words ‘a bridge between biophysics and psychology’ (Rosenblatt, 1958). To this extent, he formulated the Perceptron (Rosenblatt, 1958), which is a Hebbian-inspired computational model of learning. In this model, he showed you can ‘train’ a neuron by correcting weights between multiple inputs to one output cell. The Perceptron (Rosenblatt, 1958) is fundamental for machine learning, as a neural network is basically multiple Perceptrons structured in different layers. In current literature, these nodes are often referred to as neurons, hence ‘neural’ nets. This idea of bundles of Perceptrons feeding into a next layer should intuitively remind you of the human brain, where this is especially evident in the visual system.

The representations formed per node in neural networks also bear a resemblance to those used in the human system, specifically the visual cortex (Wang & Deng, 2021; Yamins et al., 2014). Similar to the human system, the layers in a neural network represent increasingly complex visual structure. For example, the neurons in earlier layers respond to Gabor like patches, and the deeper layers represent visual information higher in the hierarchy, such as facial features.

Apart from the mechanistic similarities between human and artificial neural systems, one other commonality is the performance of the systems. Performance has reached, and even went beyond, human-like performance in the most modern systems (Taigman et al., 2014; Wang & Deng, 2021), but only under certain conditions and with adequate pre-processing.

These examples are by no means an exhaustive list of all the components that are biologically inspired but are just mentioned as an illustration of how the human brain has influenced machine learning and why DNNs are a particular good candidate for the exploration of human face recognition. That the focus of both psychology and computer vision is on deep neural networks, is not a coincidence. Several breakthrough moments in the last decade have steered these fields in the current direction. In the following subsection, I will discuss the major contributing factors and events.

Apart from the similarities of artificial systems to the human system, some inspiration may be drawn from the field of data science in return: any model needs the selection of features that are likely to be good predictors and the model output/task must be defined. The same question should be asked by psychologists: where does input become output? For example, is distinctiveness a feature or a descriptive output? In other words, when is it part of the representational space or a decision category based on that space? Understanding the relationship between low-level and higher-level visual facial information and how that scale is related to internal representations of faces in a multidimensional space framework such as proposed by Valentine (1991) will be key to progress our understanding of face recognition. Two main questions remain: what is a facial feature/dimension, and how can we evaluate what clusters of facial variation can be considered one dimension? Although the answers to these questions seem out of reach for the time being, I believe the increasing research on the quantification of image manipulations and progression in the (neuro-)computational modelling of human face recognition will provide great insights in the following years. In the subsection ‘Current directions’ I discuss the handful of studies which already applied such an approach to better understand human face recognition.

One could also ask the question ‘Why DNNs?’ on a more fundamental level: Why should we use a computational modelling approach at all? How do we know a model is useful? This point was addressed by O’Toole and Castillo in a recent review on advances in face processing using computational approaches (O’Toole & Castillo, 2021). The authors remind the reader of advice by Box (1979), that “scientific progress comes from motivated iteration between theory and practice.” The assumption in a computational approach should not be that one model is built that can provide a conclusive answer, but that each built model informs the question that is asked and can be adapted to further define mechanisms. However, it is noted that it is important to know how a model is deviating from human processing.

7.4. Why now?

To fully appreciate the current progress in human face recognition research by studies that use DNNs, it is important to understand the major developments in deep learning (DL) that have made these techniques a useful tool to explore face recognition. Although the history of DL is

full of interactions between psychologists and computer scientists that have led to progress in tools and concepts to explore face recognition, for the sake of brevity, I will limit the discussion to go back to 2009, the year the ImageNet dataset was released (Jia Deng et al., 2009). This selection is a reflection of the biggest/most relevant contributors to the computational part of my thesis, and the field in general (as described in overviews such as: Christian, 2020, Kurenkov, 2020).

The internet

One of the major reasons the modern models perform so well is the massive number of images available to train on. However, having images is not enough, because networks need labels to train (at that time unsupervised learning was very limited). The internet provided a solution, which Professor Li Fei-Fei realised: she used the Amazon Mechanical Turk platform to have images labelled for a relatively low-cost (Christian, 2020). The first large object data base is ImageNet, published in 2009 (Jia Deng et al., 2009). Since then, multiple larger datasets have been created, allowing for the exploration of representation in DNNs.

AlexNet

As briefly mentioned above, performance of object classifiers greatly increased in 2012 when ‘AlexNet’ showed the potential of cascading multiple layers of processing units that each dealt with a different level of extraction and transformation, and won the ImageNet challenge (Krizhevsky et al., 2012). It was known, purely theoretically speaking, that a layered structured of Perceptrons could learn basically anything, with the right training data and enough layers. As mentioned above, it was very shortly before that the ImageNet dataset was released (Jia Deng et al., 2009), but the computational aspect had remained unsolved. Part of the power of AlexNet does not lie in the architecture but in execution, it is the fact that the model runs on a GPU, which allowed for much faster processing. The main reason to emphasize AlexNet as a pivotal moment, is to emphasize the timeline: it has been less than a decade since DNNs were feasible to use in academic research, it is still evolving in multiple directions and it is technically still quite challenging. This should be kept in mind as there are still many questions on how best to use this technique to explore psychological constructs and brain mechanisms, such as the aim of the studies discussed in the last section of this chapter.

New technologies

Another major contributor to the usability of DNNs in face recognition research is several advances in terms of hardware and software. Firstly, faster Graphics Processing Units (GPUs) that became available through large-scale online servers (cloud services) allowed more people to learn and develop techniques. These new technologies made it affordable for academic researchers to acquire the necessary equipment. Secondly, the move to open source and science has been a vital part of making the current research possible. For example, techniques behind AlexNet were kept open source and software, such as Google Colaboratory, and Jupyter Notebook, made it easier to share code which in turn led to a flurry of online tutorials and papers with code, making learning the techniques involved easier. Lastly, the release of TensorFlow 2.0 with the integration of Keras and TensorFlow resulted in the increased (relative) ease it is to apply these models, an influx of applications, tutorials, and additional libraries.

7.5. Current directions

As should be clear from the overview above, the use of computational models that are based specifically on DNNs in our exploration of face recognition has been quite a recent development. The integrative approach between computational modelling to understanding face recognition, and more broadly cognition, is called computational cognition. Currently, this field is more multi-disciplinary than interdisciplinary, resulting in a steep learning curve for any psychologist embarking on a journey in computational modelling. This might be one of the reasons little research using DNNs to explore face recognition has been published so far. In this section, I provide an overview of the handful of studies using DNNs currently published.

The (dis-)agreement between human and computer

As discussed in the section ‘Why DNNs’ above, it is clear that DNNs can be a useful tool to gain more insight in human face recognition (e.g. O’Toole & Castillo, 2021), but at the time, it remains unclear to what extent human behaviour and a DNN output align. Despite the many unknowns, DNNs are already being implemented practically, for example, at security

applications at the airport. In a study done recently in our lab by Hancock, Somai and Mileva (2020), it was found that DNNs do not always work in agreement with humans.

In the study (Hancock et al., 2020), the apparent race and gender of faces were manipulated and presented to DNNs. Clearly, two images showing people with a different apparent race or gender would not be identified as the same by any human observer. However, surprisingly enough, the DNNs did not pick up on the changes on these pairs, often disregarding these changes. It is difficult to say why this is the case, because commercial DNNs were used, that function as a black-box: it is unclear what the effect of the changes was on the representation as only the final match was used. However, humans and the DNNs did show agreement in terms of the pairs they found difficult, thus indicating some commonality in the underlying mechanisms. This study shows the importance of understanding the representation in the DNNs, and what specifically is considered the identity space for a DNN.

Features in DNNs

As mentioned in Chapter 1, features in psychological theory can be anything that is distinguishable in the visual information provided by a face. For example, a feature could be the distances between the eyes of a person. It can also be more abstract, such as the spatial frequency content. In Chapter 1 and 2, I discussed the reverse engineering approach of Abudarham and Yovel (2016) and the subsequent study on the exploration of critical features for face identification (Abudarham et al., 2019). In their framework, features are facial parts and a face space was concretely formulated in terms of 20 features, which were then scored by participants to assign values to each feature and create a vector per face. They used these vector values to ‘move’ faces in face space in order to explore what features are critical for identification. The authors not only tested human participants, they repeated the procedure with a DNN too, in order to explore if the features were also critical for artificial systems. They found that, similar to the human data, DNNs also show higher sensitivity to the high perceptual sensitivity (PS) changes versus the low PS changes. This means that the same features that were critical for face recognition for human observers were also critical for the DNN.

View invariance in DNNs

Another study by Abudarham, Grosbard and Yovel (2021) explored if human face recognition depends on a system that is selectively tuned to view-invariant features. To explore this question, they compared different DNNs, one optimized for objects, and the other optimized for faces. Specifically, the face-optimized DNN was trained to recognize 1000 identities, and the object-optimized DNN was trained to recognize 1000 different objects. For each of the networks the differences in facial representations between pairs of faces in different or the same view was calculated in terms of the Euclidean distances between the representational vectors (retrieved from the last fully connected layer). They found that the face-optimized DNN, and not the object-optimized DNN, did have a view-invariant facial representation deeper in the network. In the second part of the experiment, the authors explored how these findings relate to human data. However, the same task could not be used for humans because, according to the authors, humans can easily generalize across different head-views, and they argue that this would result in the same similarity scores for all different head-views. To circumvent this issue, they opted for an indirect measure of view-invariance: instead of asking participant to rate one identity in different head-views and compare those scores, participants were asked to rate similarity between two different identities who had the same head-orientation. They only presented each identity pair only in one orientation to a participant, and had multiple groups of participants for the different orientations. Thus, the human similarity ratings for the different orientations were based on different groups of participants. A similar procedure was used for the DNN, except that the DNN scored all different head-views. The authors argue that a high correlation of these distance scores across head-views is an indication of a view-invariant presentation, and that is exactly what they found.

In the second experiment in this study, which is similar to the experiment described in the section ‘Features in DNNs’, the authors explore if the critical features they identified for humans are also critical for the DNNs. Again, the face- and object-optimized DNNs were compared to each other. Their findings showed that the face-optimized DNN showed greater sensitivity to high PS changes and the object-optimized DNN did not.

Although this study provides an interesting framework to explore face representation, one can question the comparison between object- and face-optimized DNNs, because these training image sets are clearly not comparable: the face image set is much more homogenous compared

to the object image set, making view-variation more distinguishable and thus an obvious candidate for a feature. To clarify, the images of faces came from the VGGFace dataset, which presents faces in a limited range of angles, whereas the images of objects provide a wider range, but have the same number of images to train on. Currently, it isn't clear what effect this difference in training datasets may have but it should be kept in mind for the interpretation of the results.

Familiarity in DNNs

The specifics of training a network are an important part of the research design in computational cognition, especially since training could reflect becoming familiar with a face. To our knowledge, the first study that specifically attempt to explore familiarity using DNNs is by Blauch, Behrmann and Plaut (Blauch et al., 2021a), in a series of experiments that explores whether face recognition as a perceptual expertise is domain-general or domain-specific. Using DNNs, the authors evaluate to what extent knowledge of other faces are used in face recognition by examining the effects of familiarity of three types of networks: untrained, trained on objects and trained on faces. In this section, I will discuss the findings of this study, but also discuss the proposed paradigm in greater detail as it will form the background to Chapter 8 - Exploring the representation of familiar faces using deep neural networks.

Model

The architecture used by Blauch et al. (Blauch et al., 2021a) was a VGG-16 DNN (Simonyan & Zisserman, 2015) which has a total of 22 layers. 16 of these layers contain weight (hence the name, VGG-16). The layers are as follows: 2 convolutional (conv.) layers, pool, 2 conv. layers, pool, 3 conv. layers, pool, 3 conv. layers, pool, 3 conv. layers, pool, followed by 3 fully-connected layers (referred to as fc6, fc7, and fc8) in which the last fully-connected layer feeds into the SoftMax layer. A fully-connected layer is a layer that is connected with all nodes from the previous layer, and outputs to all nodes of the next layer. VGG-16 contains 3 such fully-connected layers before the Softmax layer. The SoftMax layer converts the activations from the last fully-connected layer (fc8) into probabilities per class. A class per object/identity learned is present.

Simulating pre-existing expertise

Since the main aim of the study of Blauch et al. (2021a) was to explore whether face recognition as a perceptual expertise is domain-general or domain-specific, a simulation of pre-existing expertise was necessary. They used the initial states to mimic this, and initialize 3 ways: randomly, pretrained on objects, or faces. Since the expertise hypothesis explored is not directly related to this thesis, I would like to refer for the full details on the pre-training associated with their perceptual expertise experiments to the original article (Blauch et al., 2021a). The pre-training relevant for their familiarity experiments is discussed in greater detail below, in section ‘Simulating familiarity’.

Effects of face domain experience

As expected, the authors found that the face-optimized network learned new identities better than the object-optimized and random network and was also faster in doing so. Additionally, the face-optimized network was able to recognize the newly learned identities with fewer images needed in training compared to the two other networks. Although learning slowly and performing worse than the face-optimized network, the object-optimized network performed better than the randomly initialized network. However, to reach this performance, the object-optimized network did require the maximum number of training images available. Thus, the authors conclude that experience with the idiosyncratic variability of an identity by itself is not enough, it must be accompanied with previous, extensive exposure to the item category in general.

Before fine-tuning, the face-optimized network has informative representational spaces in the deeper layers before the probability layers. The authors interpret this as evidence that even the within-identity variability of unfamiliar faces can be represented in a way that same identities will be closer than different identities and argue that “the within-identity variability learned for familiar faces allows for large improvements in verifying unfamiliar face identities.”

Additionally, the authors explored the effect of face experience in terms of learned images on the perception of unfamiliar faces and performance on newly learned faces. To simulate face experience, the network was trained on different portions of the VGGFace2 dataset; either 1%, 10% or 100% of the dataset was used. Perception of faces was defined as the performance calculated between vectors per each individual layer. The results show that the

verification performance of both familiar and unfamiliar faces increases with the increase of face experience.

Lastly, the authors explored the role of perceptual learning in the familiar face advantage by altering the depth of fine-tuning. They compared three differently fine-tuned networks, in which either all layers, only the fully-connected layers, or the probability layer was fine-tuned. In addition to the varying number of layers fine-tuned, they also varied image set and trained for objects and faces, with either 10 or 100 identities. The face-optimized networks did not show a significant effect of training depth in either (10 or 100 identities) training conditions but the object-optimized network showed significant improvement as the depth of training increased. Based on these findings, the authors conclude that perceptual learning is not necessary for the familiar face advantage, provided that there is an expert perceptual face mechanism that is created by learning from other faces.

Simulating familiarity

As this thesis is about exploring the effects of familiarity on human face recognition using (a psychological and) a computational modelling approach, the specifics of how familiarity is simulated in a DNN is of obvious interest. Blauch et al. (2021a) mimicked familiarity by performing fine-tuning of the three types of networks. Briefly explained, ‘fine-tuning’ refers to training only the last layers of a pre-trained model, whilst keeping the earlier layers frozen. To mimic familiarity, it is important that the identities that are unfamiliar were not present in the training set of the pre-trained model. In this case, all layers up to the fully-connected layers are frozen, and only the weights of fc6, fc7, fc8, and the probability layer are changed during training. Unfamiliar representations were the baseline network, before any fine-tuning took place. Familiar representations were created by training the network for 50 epochs on images of the previously unfamiliar identities, with a fixed learning rate of 0.01 and a momentum of 0.9. The network was trained on all identities in the Labeled Faces in the Wild dataset (Huang et al., 2008) that had at least 18 images. 10 of these images were removed for the test set, leaving a minimum of 8 images per identity to train on. The total number of identities, as well as the average number of images, batch size and SD are not mentioned in the paper, but if a selection of identities with a minimum of 18 images was used this left 74 identities, with an average of 44 images per person with a SD of 66 (thus, batch size remains unknown; numbers based on “All

names (with number of images for given name) as text file” file on <http://www.cs.umass.edu/lfw/#download> in December 2021).

Effects of familiarity through the network

Similar to the finding in Chapter 0, Experiment 4a, Blauch et al. (2021a) find that familiarity increases similarity. However, the effect was the strongest in the probability layer and present to a lower degree in the fully connected layer fc7. The familiarization led to greater uniformity and similarity between the non-matches in fc7, the representational space, but barely any change in the matches. The authors argue, though not completely explicit, that the effect of familiarity is similar to the function of the probability layer. As briefly discussed in ‘7.3 Why DNNs?’, the function of the Softmax transforms the activation into a distribution of probability. Each of the classes is represented by a node, and the Softmax approximates a winner-take-all. However, all the classes have some level of probability. As a result of the Softmax, similar items are closer to each other and increases the distance to other classes. However, the authors note that the latter does not seem to happen, as the non-matches show an increased similarity after familiarization. Since the effect of familiarity is the presence of more exposure to the familiar classes compared to unfamiliar (unseen) classes, it is no surprise that this effect is present in the probability layer.

A note on simulating familiarity in face recognition

A common issue in all forms of machine learning is overfitting. The system models the training data very precisely and fails to generalise well to novel examples. A manifestation of this for AFR is that the system may learn to recognise the individuals in the training set very well, but handle new identities poorly. Much of the skill in training DNNs is around prevention of such overfitting and Blauch et al. (2021a) report that they avoided it. However, the relatively poor performance of humans with unfamiliar faces suggests that we may ‘overfit’. This might make sense from an evolutionary perspective: our need is to be extremely good at recognising familiar faces. Matching two different images of an unknown individual is a very recent and artificial task. In the context of Burton et al. (2016), who argues that face variability is highly idiosyncratic and humans do not seem to learn from known identities how they should expect the variation to be in unfamiliar faces, one might argue that over-fitting is a principle necessary to explore familiarity.

Additionally, Blauch et al. (2021a) uses 50 epochs to simulate familiarity. However, it is unclear if this is a suitable training duration. As the authors show, experience, defined as number of images per identity, mattered. It is not unlikely that the number of epochs also reflect a part of experience. In a commentary following the original publication (Blauch et al., 2021b), the authors do emphasize that the goal was not to mimic human face recognition with their model, stating: “the models ... are fit to maximize performance on facial recognition, not to reproduce any details of human performance”. I will continue this discussion and return to these latter questions in the next chapter, in which I explore familiarity in representational space.

7.6. Summary and conclusions

Recently, developments in technology have resulted in a more widespread availability of the necessary hardware for deep learning, making it more accessible. For example, the creation of Google Colab allows for any user with an internet connection to run GPU-intense programs on their servers, for free. These developments have resulted in a rapid increase of open-source machine learning techniques, toolboxes and tutorials, which makes this technique very suitable for application in face perception research (and science in general).

As is hopefully clear from the literature discussed so far, DL is a tool that can be used in such an approach that allows for innovative and more comprehensive ways to interpret data and can help formulate new hypotheses. The research on familiarity in face recognition using DNNs is limited, but provides interesting directions for further research. The computational part of this thesis will revolve around the manipulation of artificial face recognition systems in order to find patterns that are similar to human behaviour. More specifically, Experiment 7 in this thesis will compare artificial and human face recognition.

8. Chapter 8 - Exploring the representation of familiar faces using deep neural networks.

Chapter main message

A method of familiarization of faces in the form of fine-tuning of a pre-trained DNN is used to explore the two main findings presented in the thesis: familiar faces are perceived as more similar overall than unfamiliar faces (Experiment 4a), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). Additionally, generalizability was explored as a marker of successful human-like familiarization. In the first experiment (7a), FaceNet was used, due to the high performance of the network and ease of implementation. No effects of familiarity were found. In the second experiment, the same procedure was repeated with VGGFace (Parkhi et al., 2015) as used in Blauch et al. (2021), to confirm the findings of Experiment 7a. In Experiment 7b, a main effect for Transformation Level, but not for Familiarity, was found. Additionally, there was a significant interaction between the two for the two largest Transformation Levels. In the third experiment, the model in Experiment 7a was retrained without any frozen layers to achieve maximum performance on the test set. Similar to Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level was found in Experiment 7c, there was only a main effect of Transformation Level. The findings of these three experiments highlight that the parameters used in training of networks might be as important as the architecture. Additionally, the findings of Experiment 7b indicate that the effects of familiarity on sensitivity to change can be found in the representation of a face. Lastly, generalizability does not appear to be a marker of human-like familiarization.

8.1. Introduction

As discussed at great length in this thesis so far, it is well known within face research and beyond that people are better at processing familiar than unfamiliar faces (e.g. R. J. Johnston & Edmonds, 2009), and although there is an abundance of research on the effect of familiarity, it remains unclear what underlies this enhanced performance. As discussed previously, a common proposal is that the representation and processing of familiar faces are different from that of unfamiliar faces (R. J. Johnston & Edmonds, 2009; Ramon & Gobbini, 2018). Recent work from

Ritchie & Burton (2017) shows that variability is important for learning a new face. Additionally, the variability within images of a specific identity appears to be idiosyncratic information that is key for learning new faces, as it becomes part of the representation of a familiar face (Burton et al., 2016). These findings indicate that familiar faces have a richer internal representation than unfamiliar faces. However, what ‘richer’ means is unknown, though the variability will most likely play a role in this. In other words, it remains unclear what these differences in internal representations specifically are.

In the first part of this thesis, I have set out to gain a deeper understanding of the difference between familiar and unfamiliar faces by exploring the effect of familiarity on similarity ratings. To recap, in Experiment 1 and 2, I found no significant effect of familiarity in a task that asked participants to match an image of a person by altering lip and eye brow thickness (high perceptual sensitivity), or distance between the eyes and the eye brows and the width of the mouth (low perceptual sensitivity). To ensure these results would generalize to other features, the influence of familiarity on the relationship between image-based spaces and perceptual similarity of multidimensional transformations was explored in Experiment 3 and 4a (Chapter 3, Somai & Hancock, 2022). In that study, I used multidimensional image transformations based on PCA to transform faces. Famous averaged faces were used with 2, 3, 5 or 7 principal components transformed. After a familiarity check, participants were asked to rate the similarity of the original and the transformed image. The images were shown with a delay: the original was shown, followed by a mask, followed by the transformed image, followed by another mask and lastly the question ‘Could these be two pictures of the same person?’. Responses were given on a six-point scale: “1 – certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. The results showed that familiar faces are rated as more similar than unfamiliar faces.

This result can feel counterintuitive at first glance, because it is known that people are better at matching faces of familiar faces as opposed to unfamiliar people, so you could also expect that the familiar faces look more dissimilar because it would be easier to spot differences. A possible explanation of the findings of Experiment 4a is that people focus on finding similarity instead of focusing on dissimilarity, and there might be more similarity in a familiar face because it has a richer representation than an unfamiliar face. In this case, if the same number of components are changed for familiar and unfamiliar faces, the effect of the change might be

perceived as smaller for familiar faces than for unfamiliar faces because the components that remain after the change are more informative for familiar faces than for unfamiliar faces. In other words, the relative change for familiar faces is smaller. In turn, the smaller relative change in familiar faces will lead a higher similarity rating compared to unfamiliar faces. Thus, the idea of relative change is based on the assumption of a richer internal representation of familiar faces compared to unfamiliar faces.

In Experiment 5, I focussed on dissimilarity opposed to similarity and found that familiarity increases sensitivity to change, but only at the largest change in the experiment. This finding would fit with the idea of a ‘richer’ representation for familiar versus unfamiliar faces, in which a familiar face has a clearer identity boundary in face space than unfamiliar faces (similar as described in Beale & Keil, 1995). To further explore this hypothesis, I will explore the effect of familiarity in a deep neural network (DNN).

As discussed in the previous chapter, Blauch et al. (2021a) used the fine tuning of a pre-trained net to simulate familiarity. Similar to the finding in Chapter 0, Experiment 4a, Blauch et al. (2021a) finds that familiarity increases similarity. However, the found effect was the strongest in the probability layer and present to a lower degree in the fully connected layers, fc6 and fc7. The familiarization led to greater uniformity and similarity between the non-matches in fc7, the representational space, but barely any change in the matches. One major difference between the finding of Blauch et al. (2021a) and the behavioural findings discussed in this thesis so far, is that I am exploring the effect of well-controlled artificial changes as opposed to the effect of within-identity image variation. Although the work of Blauch et al. (2021a) is informative, it is difficult to generalize their findings in the context of sensitivity to change and how this might be expressed in representational space.

In this chapter, I explore the effect of the previously used well-controlled PCA transformations on representational space. Specifically, I will compare the size of the difference between face representations. Since the vector length of the representational layer is held consistent over all images, a richer representation would translate into a more robust – i.e. consistent – vector. This in turn means the distance between two vectors of the same identity (depicted on different images) will be smaller for familiar than unfamiliar faces.

8.2. Methods Experiment 7a

For the current study, a pre-trained FaceNet (Experiment 7a and 7c) and VGGFace (Experiment 7b) will be used to quantify changes to an image of a face. Familiarity will be mimicked by tailoring the last layers to a set of identities, i.e. fine tuning. ‘Fine-tuning’ refers to training only the higher-level layers of a pre-trained model, whilst keeping the earlier layers frozen. After this, I will compare the vectors of the last fully connected layers (here: batch normalization layers) of pairs of unfamiliar or familiar faces and see if there are differences in similarity. In the current chapter, I will consider only Euclidean distance as that is the measure commonly used in (psychological) face literature. In the next chapter, I explore three other distance measures that are commonly used in machine learning to see how they compare to one another and to explore different mechanistic explanations for the behavioural data presented in Chapter 0 and 5.

Image sets and transformations

A selection of identities in our PCA set was made based on the identities that were present in the VGG set 2 which was used for training this particular pre-trained FaceNet. Any identities that were present in both the VGG face set 2 and our face set were removed, this was checked manually for all the identities present in our PCA. Even when using a pre-trained model, a relatively large number of images is needed for fine-tuning. Based on a previous run using a DeepFace architecture and the number of nodes in fc7 (4096), it is estimated that 171 images per identities for training (4096 divided by 24) are needed. DeepFace was used in piloting because it is an easy to implement model with a low training time. The depth of the model is limited, with a total of 8 layers. Depth of DNNs is related to performance, thus, this model served to find the absolute minimum number of training images to show effects of fine-tuning. I.e. a deeper network will have more hidden layers that can help represent the classes, thus needing less images to train on, if a pre-trained model is used. This should translate to a better performance of deeper networks compared to a network with less layers, such as DeepFace. Although DeepFace was not used in the final experiment, the above chance performance of this lightweight model was used as an indication of sufficient images for learning.

After removal of the identities present in the original training set, 49 identities remained. One additional identity was removed in order to have a balanced set with 24 identities in each Familiarity condition. Thus, this left a total of 48 identities which the network was initially

unfamiliar with. For each of the 48 identities, an image set of 171 images was created. Firstly, using the software ‘Bulk Image Downloader’, images were retrieved from Google. Secondly, using ‘Visipics’ the duplicates were removed (with the larger image always kept). All images were then converted to the same file type using ‘IrfanView’ and ‘AF5’. Then, all images were cropped and aligned using a Matlab script. Any images with a height smaller than 200 pixels were considered too small and discarded. Lastly, using the DNN provided by the University of Surrey (which is not publicly available unfortunately, access for this project was gained through the FACER2VM project), a ranking based on similarity was made and the top 171 most similar looking images were selected for the image set of that identity. This step is necessary because the images were downloaded in bulk, and multiple identities might be shown on the image, that would have then be cropped out to a single image. By selecting the top 171 images, it was ensured that only the relevant identity was selected. The similarity was based on similarity to the first image present in the image set for that identity. Thus, the first image for every identity was chosen to be an image that serves as a good reference for that identity.

Stimuli preparation and creation for the images used in the task follows the same procedure as Somai & Hancock (2022), to minimize differences in design between the computational and behavioural design in order to keep the comparison as clean as possible. Face image stimuli were generated from a PCA-based face model, written in Matlab. The PCA space was calculated from 72 eye-aligned images, each an average of 10 images of a famous (mostly Caucasian) female. Use of average images helps to reduce variations due to lighting and viewpoint. Even so, the first few components encode image properties such as lighting and head movements, so components 10 or higher were selected to alter. The SD for each component was scaled according to the variance in the (72- eye-aligned) original images and the size of the transformation was 4 SD. The total stimulus set consisted of the original images of 48 identities and transformed images that had either 2, 3, 5 or 7 colour components (components 10 – 11, 10 – 12, 10-14 and 10 – 16 respectively) in the calculated PCA space changed. Thus, this is a total of 196 pairs: 4 transformation sizes for both 24 familiar and 24 unfamiliar faces. An example of the transformation is shown in Figure 46. In the classifying task the original and the positive transformation were used (which are the transformations that look less distorted compared to transformations from the negative end).

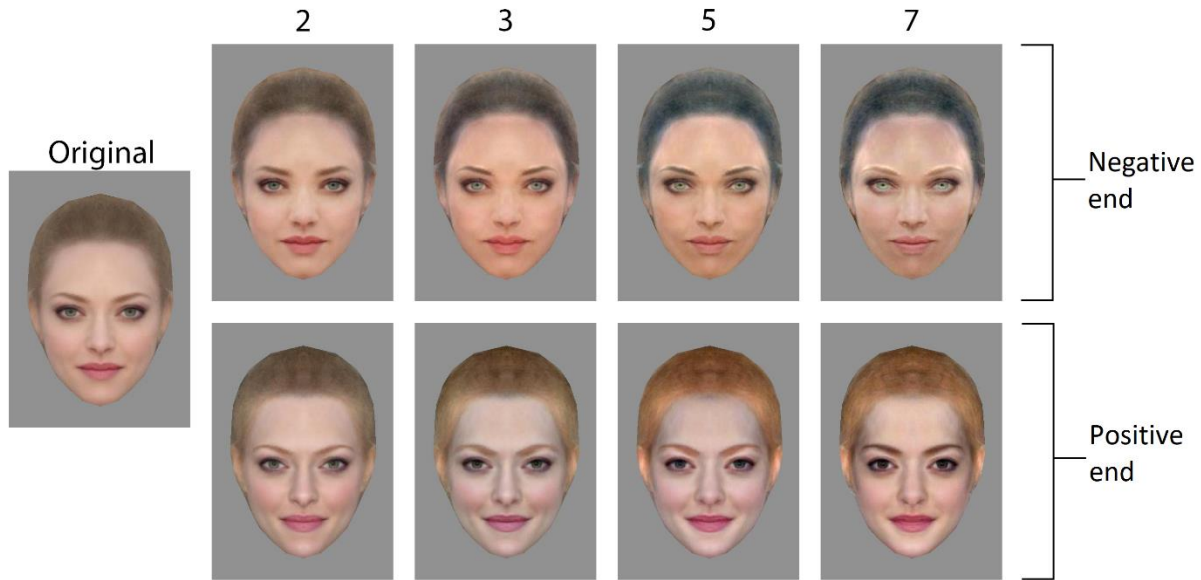


Figure 46 Example of the transformations.

General procedure

All images are fed through the model in order to generate an image-vector from the representational layer. Then for all pairs (a pair of an original image and one image with either 2, 3, 5 or 7 components changed) the difference between the vectors is calculated using Euclidean distance.

Pre-processing (setting up the testing and training data)

A random selection of 24 identities out of the 48 is selected to divide the stimuli into 2 sets. To clarify, 24 identities are selected from the 48 available identities and will form 1 subset. The 24 remaining identities form a second subset. The following procedure follows the image pre-processing mentioned above. All images are ranked by their DNN match score and the top 171 images are selected. 20% of the images need to be separated to form the test set, which are 34 images. To ensure that the images in the training and test set have the same similarity rating on average, it is important to pick images from throughout the 171 ranked images. To this end, the ranked 171 images are divided in 34 bins and 1 image is taken out of each bin and placed in the testing folder. So, each bin will contain 5 images, and 1 of those 5 will be placed in the test set,

while the remaining 4 go into the training set. All remaining ones are placed in the training folder. This division in bins was used to avoid any major differences in DNN match scores between the testing and training folder.

To avoid any effects specific to stimuli set, the entire procedure was repeated from scratch with the 24 identities previously untrained used for the fine-tuning. In other words, the procedure was counterbalanced by training both halves of the data set, so that all identities were familiar in one run and unfamiliar in another. To avoid any effects of the way the identities were divided into two halves, there were 15 different ways the identities were split in two. Thus, the complete procedure was run 30 times (for an overview, please see Figure 47).

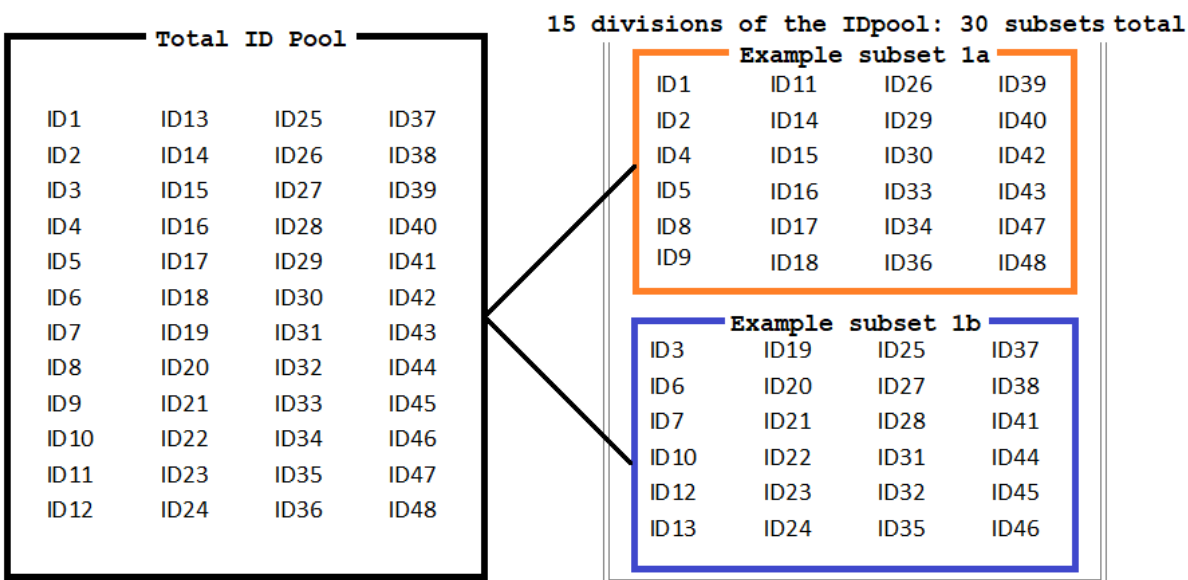


Figure 47 Example of creating one pair of subsets.

Training the model: creating 'familiar' faces

In the current study, I chose a pre-trained version of FaceNet as found in the DeepFace library. The specific version of FaceNet was trained on the VGG face 2 set (REF from PC). The architecture of FaceNet is based on inception_resnet_v1.

There are 48 identities in the total identity set. As mentioned above, to mimic familiarity 24 of these identities are used in the training to make them familiar. The 106 higher level (last)

layers of a pre-trained FaceNet was trained on a set of 24 identities, which is called fine tuning. The leftover 24 identities were used as unfamiliar faces.

To find successful hyper parameters (i.e., training parameters), I set the goal that models need to converge (reach a loss smaller than 1) and reach an accuracy of at least 70% during training. Due to a limited time budget, a random search was employed to find the hyper parameters. Initially, hyper parameters were estimated on one image set. Once the search space was smaller, hyper parameters were tweaked, based on several image sets. Once the hyper parameters had been set, these image sets were deleted and replaced by new ones.

The pre-trained model is loaded without a classification layer, thus for training a Softmax layer with 24 classes was added. This model has 447 layers in total, of which the first 301 layers are frozen and the last 106 are used to train. This model has a total of 22,808,144 parameters of which 9,829,504 are trained. The number of epochs used in training was 150 and batch size was 16. A learning scheduler was used that started after 5 epochs with a base learning rate of 0.015 and ended at 45 epochs with a final learning rate of 0.001. In case an accuracy value below .70 occurred on the test set and/or the model did not converge (i.e. did not reach a loss smaller than 1 during training), training was repeated. In case the value of the accuracy remained under .70, the subset was replaced with another (random) division.

The performance of the models had to be a minimum of .70 on the evaluation to avoid difference in task performance of the DNN on the face matching task caused by accuracy of the model. This level of performance might seem low, but has been set to compromise for a small training set. Additionally, it fits the goal of the experiment: the goal is comparing two images of the same identity to each other in which one of the images has a small change and explore if familiarity can be simulated by training. High performing DNNs are known for their ability to generalize learned features to unseen, thus ‘unfamiliar’, images, which interferes with the goal of the experiment. Thus, although it’s important to have a performance above chance as proof that it has learned, I expect the effects I am exploring will be found in a model that performs below ceiling.

The final, Softmax layer performs classification and is needed during training. During test, it would provide an identification for an input face. However, I’m interested in the representation of unfamiliar and familiar faces, thus the image classification itself is not relevant but the activation within the network is what represent the face. Thus, the SoftMax (last) layer is

not relevant, because that consists of the classes (identities in the case of faces) and that is not informative about what features are activated. For this reason, the SoftMax is removed after training and the calculation of the Euclidean distance is performed on the last fully connected layer (in this network the batch normalization layer). This representational vector consists of 128 values.

Face matching task

The exact same task as in Chapter 0 Experiment 4 (Somai & Hancock, 2022) was used in which two images of an identity were presented sequentially to the model, one original image and one image that had a number of components changed. The model returned a 128-long vector. As an equivalent to the similarity score in humans, the Euclidean distance was calculated between the vectors of each pair.

Data analysis

To explore the effects of Familiarity, the same linear mixed effects model was used as in the human behavioural experiment, with the similarity rating as dependent variable. The model included Familiarity and Transformation Level as categorical factors (Familiarity: familiar and unfamiliar coded as -0.5 and 0.5 respectively to allow for the analysis of the interaction; Transformation Level: 2, 3, 5, or 7), and the intercept, slope and interaction of Familiarity and Transformation Level in the performance per model as a random effect. Due to convergence issues, the main effects were excluded from the random effects. To explore the main effect of Familiarity, a similar model without the interaction effects included was used. The parameter estimate (b), standard error (SE), t value and p value are reported. The linear mixed effect analysis was performed in R (R Core Team, 2019) with *lmerTest* (Kuznetsova et al., 2017). The threshold for significance was set at $\alpha = 0.05$. Normality of the residuals on the model were inspected visually and showed no violation.

8.3. Results Experiment 7a

The descriptive statistics of experiment 7a are shown in Figure 48. On average, the models had an accuracy of 1 on the last epoch of training, $SD = 7.44E-04$ and an accuracy of .747 on the evaluation set, $SD = .015$. The results show that increasing the number of components transformed decreased the similarity rating for all levels compared to the base change (2 components changed, all p-values are $p < 0.001$. Secondly, no significant effect of Familiarity (estimate = -0.01, $SE = 0.02$, $t = -.30$, $p = .769$) was found and there was no interaction between Familiarity and Transformation Level (see Table 15 - Table 17 in Supplementary materials – Chapter 8 for p-values per Transformation Level).

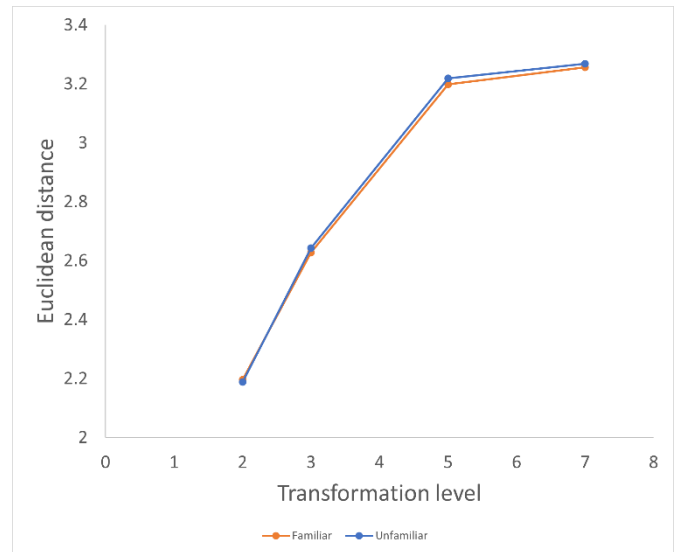


Figure 48 Euclidean distance per Transformation Level

8.4. Discussion Experiment 7a

In this experiment, I explored whether familiarization would lead to a difference in the representational space of the DNN. Familiarization was achieved by fine-tuning a pre-trained FaceNet to half of the identities available and the other half remained unseen, to mimic ‘familiar’ and ‘unfamiliar’ faces. The findings indicate that this is not the case, which is in line with the work of Abudarham and Yovel (2018), in which they found that the same features are used for the identification of familiar and unfamiliar faces.

However, before any such conclusion can be reached, more evidence is needed that fine-tuning in the manner applied here is a valid way of simulating familiarity. After fine-tuning on a relatively small set of images, performance of these models was on average .747 on the test set. Above, I reasoned that it is important to have a performance above chance as proof that it has learned but that I expected the effects I am exploring will be found in a model that performs below ceiling. Especially because the experiment is focused on finding a relative change, between an original and transformed image, I assumed the performance (target performance and

achieved performance, see Table 22 in Supplementary materials – Chapter 8) fitted the goal of the experiment. These assumptions can still be true, but in order to validate the result, the experiment needs to be reproduced using another architecture.

Additionally, the findings of Experiment 7a contradict those described in Blauch et al. (2021a). As discussed in Chapter 7, Blauch et al (2021a) found that familiarity increases similarity, although the found effect was the strongest in the probability layer and present to a lower degree in the fully connected layers, fc6 and fc7. The contradiction between their findings and those reported here could be an indication that the findings in the current experiment, or their experiment, are architecture specific.

Alternatively, the specific way the network was trained could have been insufficient to achieve an effect of Familiarity. One could argue a ‘good’ DNN is one that has no familiarization effect as it will generalize its matching abilities to all images shown. Fine-tuning should logically have the familiarization effect, as the layers are literally tuned to increase performance on a small set of identities. The lack of an effect of Familiarity could be an indication that the layers selected for training could not adequately use the representations from the pre-trained (frozen) layers to build new representations. To further explore if familiarization would lead to a difference in the representational space of the DNN, the Experiment is repeated with a different architecture.

8.5 Experiment 7b

Since the findings of Experiment 7a contradict the findings described in Blauch et al. (2021a), the experiment was repeated with an architecture that more closely resembles the one used in the paper. To this end, I choose a pre-trained version of VGGFace as found in the DeepFace library. As

8.6. Methods Experiment 7b

Image sets

In this iteration of the experiment, VGGFace is used. The pre-trained model I obtained was pre-trained on the original VGG Face set, not VGG face set 2. This means that all the steps to create the image sets that were undertaken in Experiment 7a need to be repeated because the model

should be untrained on any of the identities used in this experiment. Thus, the presence of the identities was once again manually checked. To allow for a better comparison between the experiment, a set of 48 identities was created.

Model and training

In the current study, I choose a pre-trained version of VGGFace. This specific version of VGGFace was trained on the (first) VGG face set (Parkhi et al., 2015). A selection of identities in our PCA set was made based on the identities that were absent in the VGG set (which was used for training VGGFace).

The architecture is loaded without a classification layer, thus for training a Softmax layer with 24 classes was added. This model has 38 layers in total, of which the first 30 layers are frozen and the last 8 are used to train. This model has a total of 145,002,878 parameters of which 130,288,190 are trained. The number of epochs used in training was 125 and batch size was 16. Again, performance of the models had to be a minimum of .70 on the evaluation to avoid difference in task performance of the DNN on the face matching task caused by accuracy of the model, and had to converge during training. Similar to Experiment 7a, the SoftMax is removed after training and the calculation of the Euclidean distance is performed on the last layer before the Softmax, which is a Flatten layer. This representational vector consists of 2622 values.

8.7. Results Experiment 7b

The descriptive statistics of Experiment 7b are shown in Figure 49. On average, the models had an accuracy of .999 on the last epoch of training, SD = $6.24e-04$ and an accuracy of .994 on the evaluation set, SD =

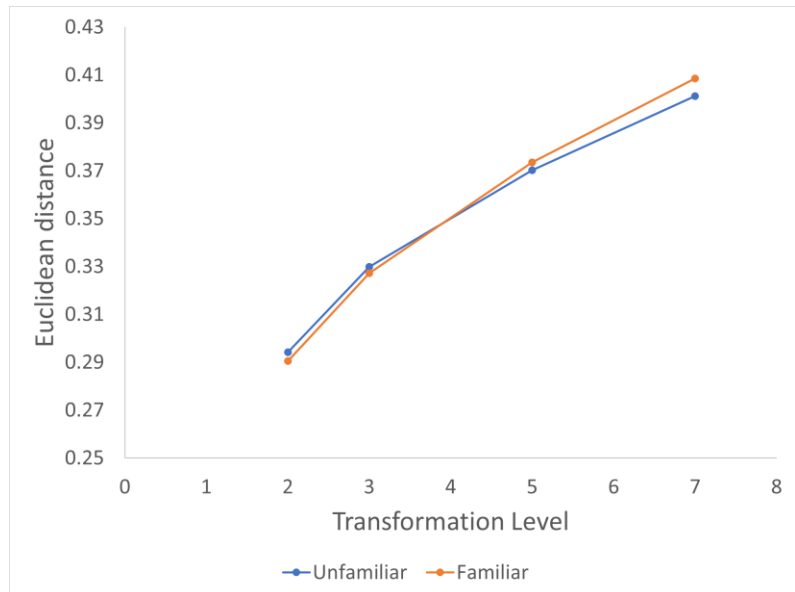


Figure 49 Euclidean distance per Transformation Level

1.57e-03. A similar LMM as Experiment 7a was used. However, since there were no

convergence issues, the main effects were included in the random effects. The results show that increasing the number of components transformed decreased the similarity rating for all levels compared to the base change (2 components changed, all p-values are $p < 0.001$). Secondly, no significant effect of Familiarity (estimate = $1.07e-03$, SE = $2.43e-03$, $t = .44$, $p = .288$) was found, but there was a significant interaction found between Familiarity and Transformation Level for Transformation Level 7 (estimate = $1.09e-02$, SE = $4.29e-03$, $t = 2.55$, $p = .063$; see Table 18 in Supplementary materials – Chapter 8 for all p-values per Transformation Level).

8.8. Discussion Experiment 7b

In this experiment, I repeated Experiment 7a with another architecture, VGGFace, to further explore if familiarization would affect the representational space of the DNN. A main effect for Transformation Level was found, but not for Familiarity. However, there was a significant interaction between the two for the largest 2 transformation levels. These findings are in line with Blaich et al. (2021a). Although the difference in absolute value between familiar and unfamiliar is small, the pattern of differences partially mimics one of the main findings presented in this thesis: increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b).

Comparing the average performance of these networks, it becomes immediately apparent that VGGFace (Experiment 7b) has a smaller difference in model accuracy between training and test data than FaceNet (see Table 22 - Table 23). In other words, the learned face matching ability of VGGFace appears to generalize better to unseen images. One could argue that this generalization could be considered a marker of familiarity. As there currently is no strict definition of familiarization in computational human face recognition approaches, this question is worth exploring further.

In Experiment 7a, it proved to be difficult to minimise the loss of accuracy between training and test data when using fine tuning. As discussed in the previous chapter, Blaich et al. (2021a) used the fine tuning of a pre-trained net to simulate familiarity. The use of fine tuning has several reasons. Firstly, there is the assumption that the early layers in a network reflect the low-level visual system (Nonaka et al., 2021). Thus, freezing the layers that reflect low level visual features should have little effect when training a network to learn faces, and these initial layers should contain only features necessary to represent other similar dataset when large

datasets are used for training (such as the VGGFace2 dataset in this case). Secondly, there are several practical limitations that make fine tuning the most suitable approach, such as a relatively small set of images to train on.

The lack of an effect of Familiarity in FaceNet could be an indication that the assumption that the early layers in a network reflect the low visual system is not applicable for this network and that it therefore struggles to learn new identities through face learning using fine-tuning. If this were true, training without any layers frozen, should lead to a better generalization of performance. This idea was piloted and was indeed true. To gain more insight into familiarization of DNNs, Experiment 7a was repeated with a different training method.

8.9. Experiment 7c

In this experiment, the aim is to gain a better understanding of training methods to mimic familiarity. Additionally, the findings of Experiment 7a and 7c will be compared in order to assess if generalizability – defined as the difference accuracy of the models between training and test set – can be used as a marker of successful (human-like) familiarization. If so, Experiment 7c should show similar familiarization effects as found in Experiment 7b.

8.10. Methods Experiment 7c

All methods and procedures are the same as Experiment 7a, bar the number of layers frozen. In this experiment, none of the layers are frozen, i.e., all are trained on the new identities.

8.11. Results Experiment 7c

The descriptive statistics of experiment 7c are shown in Figure 50. On average, the models had an accuracy of 1.000 on the last epoch of training, $SD = .015$ and an accuracy of .997 on the evaluation set, $SD = .0002$ (see Table 24). The results show that increasing the number of components transformed decreased the similarity rating for all levels compared to the base change (2 components changed, all p-values are $p < 0.001$). Secondly, there was no

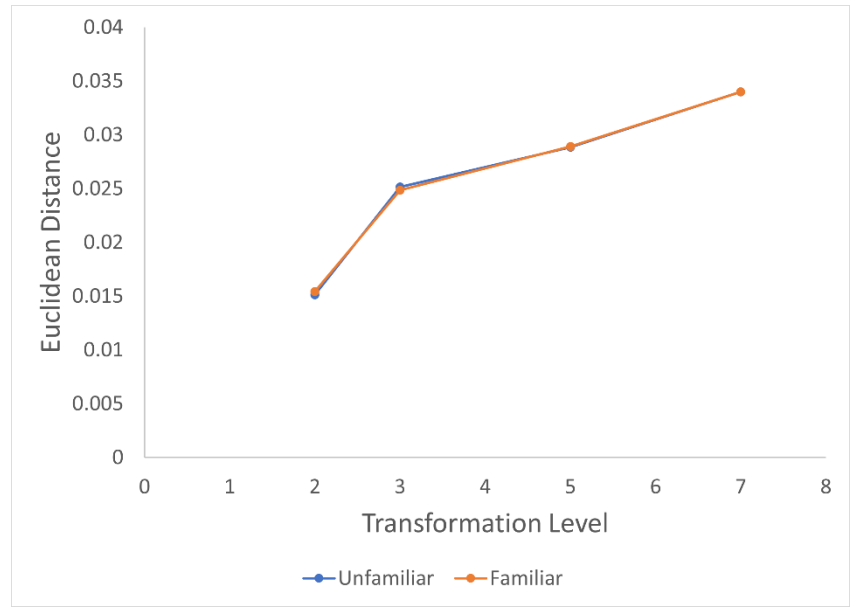


Figure 50 Euclidean distance per Transformation Level

significant effect of Familiarity and no interaction found between Familiarity and Transformation Level (see Table 19 - Table 21 in Supplementary materials – Chapter 8 for p-values per Transformation Level).

8.12. General discussion and evaluation of this framework

In this study, I explored if familiarization would lead to a difference in the representational space of the DNN in order to gain deeper insight into two main findings presented in this thesis: familiar faces are perceived as more similar (Experiment 4a), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). Additionally, generalizability was explored as a marker of successful human-like familiarization.

Familiarization was achieved by fine-tuning a pre-trained FaceNet (Experiment 7a) and VGGFace (Experiment 7b) to half of the identities available while the other half remained unseen, to mimic familiarity. Additionally, FaceNet was trained with loaded parameters but no frozen layers (Experiment 7c). In Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found, there was only a main effect of Transformation Level. In Experiment 7b, a main effect for Transformation Level was found. Additionally, there was a significant interaction between the two. In Experiment 7c,

similar to Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found, there was only a main effect of Transformation Level.

The findings of Experiment 7b, are partially in line with two main findings presented in this thesis: the significant interaction between Familiarity and Transformation Level are representative of the findings in Experiment 5a and 5b, in which the results show that the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces. Thus, it appears that the effects of familiarity on sensitivity to change can be found in the representation of a face.

That the representation of familiar faces is different from unfamiliar faces might appear to oppose the findings of Abudarham and Yovel (2018) that the same features that were critical for face recognition for human observers were also critical for the DNN. However, I believe it compliments their findings. In their framework, features are facial parts and a face space was concretely formulated in terms of 20 features, which were then scored by participants to assign values to each feature and create a vector per face. They used these vector values to ‘move’ faces in face space in order to explore what features are critical for identification. They found higher sensitivity to the high perceptual sensitivity (PS) changes versus the low PS changes. Thus, their question is qualitative in nature: which dimensions in face space are important and what is used in face matching tasks. The question at hand in the current study, however, is quantifying the difference between familiar and unfamiliar faces: even if faces are on the same dimensions, do you have a more accurate representation of familiar versus unfamiliar faces?

One could be tempted to see the lack of a significant main effect of Familiarity as support for the findings of Abudarham and Yovel (2018), however, this is not correct. As stated above, their question is which dimensions in face space are important and what is used in face matching tasks. This is very different from how DNNs are set up: DNNs are trained on extremely large data sets so as to create all the dimensions needed to properly represent a face. Thus, the lack of main effect of Familiarity was to be expected, although it remains unclear if different training and different architectures could show a main effect of Familiarity. I will return to this point in the general discussion in Chapter 10.

As mentioned in the discussion of Experiment 7a, the findings of Experiment 7a contradict those described in Blauch et al. (2021a) and those found in Experiment 7b, which could be an indication that the findings in this experiment are architecture specific. These

findings were reproduced using a different training approach. A lack of significant findings should always be treated with caution in terms of drawing conclusions. However, based on the combined findings of Experiment 7a and 7c, the most obvious explanation would be that FaceNet is an architecture with a non-brain-like structure. Two recent papers have suggested a ‘Brain-score’ for artificial neural networks (ANNs), in order to assess how similar the activity within layers of the ANN is to that of measured brain activity. Note that the emphasis in these studies is the brain, not behaviour. Comparisons are made between ANNs and encoded and/or decoded brain activity. A more in-depth discussion of these papers can be found in the general discussion, in ‘Brain-likeness paradigms & biological plausibility’, but it is worth noting that VGG16 – on which VGG-Face is based - scores relatively high on both the brain score of Schrimpf et al. (2020) and Nonaka et al. (2021), .521 and .38 respectively. Unfortunately, these scores are not (yet) available for FaceNet, which is based on Inception-ResNetv1.

Apart from the difficulty in interpreting the output of the DNNs in the absence of knowledge on their brain-likeness, another note that can be made about this paradigm is the definition of familiar. As discussed in Chapter 1, section ‘1.3. What is familiarity?’ the definition of what a ‘familiar’ face is, is not as straightforward. Without a concrete definition, it will be difficult to mimic the equivalent level of ‘familiarity’ in a DNN. However, as discussed to a greater extent in Chapter 7, the assumption in a computational approach should not be that one model is built that can provide a conclusive answer, but that each built model informs the question that is asked and can be adapted to further define mechanisms. Thus, a different training procedure should not mean that the model cannot be informative, but it is important to know how a model is deviating from human processing. I will return to this point in the general discussion, in ‘Brain-likeness paradigms & biological plausibility’.

In these experiments, instead of focussing on architectural variations or different tasks, I use an artificial face recognition system that performs extremely well and attempt to manipulate the training to get more human-like responses. Although this succeeded to the extent that a similar pattern of response can be seen in the DNN’s representation vector (Experiment 7b), the scale of the response remains non-human-like. The differences in the DNN’s representational vectors were small enough that identities are most likely to still be matched as the same identity. Additionally, one can note the focus on only the representational layer before the classification layer in these experiments as a limitation. As discussed in Chapter 7, section 7.3, different layers

in DNNs contain different information. Similar to the human system, the layers in a neural network represent increasingly complex visual structure. For example, the neurons in earlier layers respond to Gabor like patches, and the deeper layers represent visual information higher in the hierarchy, such as facial features. The output layer only represents the label identifying the face, whereas the earlier layers represent features that are useful in discriminating among faces. Therefore, this approach was chosen because this thesis is focussed on the representation of faces, the image classification itself was not relevant but the activation within the network is what represents the face. Thus, the findings are meant as an initial indication that this approach could yield useful insights that can later be expanded to other aspects of the DNNs. This study provides an example of how hypotheses and behavioural data can inform DNN parameters in an aspect that has been understudied so far: training.

Lastly, I explored whether generalizability – defined as the difference in accuracy of the models between training and test set – can be used as a marker of successful (human-like) familiarization. In Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found, there was only a main effect of Transformation Level. However, there was a drop in accuracy between training and test: on average, the models had an accuracy of 1 on the last epoch of training, $SD = 7.44E-04$ and an accuracy of .747 on the evaluation set, $SD = .015$. The lack of an effect of Familiarity in FaceNet in Experiment 7a could be an indication that the training was insufficient through face learning using fine-tuning. If this were true, training without any layers frozen (since a pre-trained network was used), should lead to a better performance and generalization of performance between training and test. Although this was true, the findings of Experiment 7c were similar to Experiment 7a: no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found, there was only a main effect of Transformation Level. Thus, it appears generalizability is not an indication of a human-like familiarization.

In sum, this study shows that the effect of familiarity appears in the representation of a face in a VGG-16 architecture, and generalizability does not appear to be an indication that familiarization in a human-like way has been successful. These findings can serve as a base when considering hyperparameters to train a network to mimic human familiarity. For example, when training with small datasets, it can be beneficial to use fine tuning. But as we have seen in this study, there is a relationship between architecture and training that remains unclear for now,

although it is likely that ‘brain-likeness’ plays a role. Additionally, the findings of Experiment 7b indicate that the effects of familiarity on sensitivity to change can be found in the representation of a face, and this paradigm provides novel ways to explore the representation of faces further. This study outlines a framework how a computational approach can provide insight into behavioural findings.

Evaluating the use of a computational approach to explore effects of familiarity on human face recognition: challenges and limitations

Although the work in this chapter has been presented as a straightforward procedure, the approach presented in this chapter is a departure from the existing computational approach that is used to explore human face recognition. The main difference with existing approaches is the concept of one model to reflect one observer, leading to the training in 30 models instead of comparing human data to just one trained network. In the following section, I will discuss several practical challenges and limitations one faces when employing such approach.

The first challenge of this approach is the selection of a network. Preferably, this would be inspired by the literature on brain-like models (e.g. Nonaka et al., 2021; Schrimpf, Kubilius, Hong, et al., 2020; Schrimpf, Kubilius, Lee, et al., 2020) which I will discuss further in the general discussion. However, due to limited processing power and a limited time budget, it was necessary to use pre-trained models for this project. Although there are increasingly more pre-trained models available, detailed information on what they have been originally trained on is often missing or incomplete. This does not per se have to be an issue to explore human face recognition, however, is essential information when exploring familiarity (and other identity-related topics, such as the other-race effect and face diet).

The second major challenge of this approach is setting criteria to define the hyper parameters. Similar to the process of ‘business understanding’ in the CRISP-DM model (Shearer, 2000), there is a necessity to define what criteria are needed for a model to be suitable to address the question that is being asked. However, there are currently no clear guidelines to set these in a way that it increases the likelihood of a human-like performance. Even the definition of human-like performance is challenging, as there is no behavioural data available on my particular image sets. However, the literature reviewed in Chapter 1, in particular the section 1.5. ‘Facial variation’ and ‘Two ends of a scale – measuring individual differences, deficits and enhanced performance’

can inform a threshold. In formulating the criteria I have considered human performance, for familiar and unfamiliar, but also the practical limitations of the approach, which I will discuss below. The images used for training can be rather challenging to match (especially in their cropped variant), and the aim of this work is meant as a starting point, thus 70% accuracy should have been sufficient. However, it should be clear that these criteria should be seen as parameter that needs to be considered in future research, as it will undoubtedly affect the human-likeness of models.

The third major challenge is finding the ‘right’ number of layers to train. Once an architecture that has pre-trained models, enough code available for implementation, and detailed information of the training set has been identified, the next question is defining the fine-tuning parameters. In my case, I have set the criteria for the right amount to centre on the loss and accuracy of the model: models needed to converge (reach a loss smaller than 1) and reach an accuracy of at least 70% during training. Apart from the above-mentioned theoretical difficulty in defining these criteria, the process of finding the fine-tuning parameters to meet the criteria is a laborious and time-intensive procedure. The deeper the network, the longer the process is to find the ‘right’ number of layers to train. In the current study, due to a limited time budget, a random search was employed to find the hyper parameters. Initially, hyper parameters were estimated on 1 image set. Once the search space was smaller, hyper parameters were tweaked, based on several image sets. Once the hyper parameters have been set, these image sets were deleted and replaced by new ones. The ‘right’ number of layers will be the number of layers that meets the criteria, but as discussed before, this too can be varied. Due to the high-dimensional nature of DNNs, there is no guarantee that these hyper parameters work on the first attempt of training. In these experiment, 25-35% of training runs had to be repeated because the model did not meet the criteria.

Lastly, the fourth major challenge – but the most prominent challenge in my PhD – is the technical skill required to implement new findings in computational neuroscience and techniques in computing science. Most, if not all, of the challenges mentioned so far has or will have a more time-efficient solution available in the broader landscape of computing science. However, the implementation of such techniques requires specific expertise, which can be difficult to identify for researchers who are not submerged in the rapidly developing field of computing science.

In summary, the framework presented in this chapter is a computational approach that allows the exploration of human face recognition by a direct comparison to a DNN on the same task. The use of defining training procedures that can be used for multiple training sets is laborious and time intensive. Some of these challenges will diminish over time, such as the lack of detailed information on training procedures, and – hopefully in the near future – a limited amount of processing power. Additionally, as computational approaches become more common in the exploration of human face recognition, more literature will be available that can constrain the search space when exploring hyper parameters.

9. Chapter 9 - The relationship between distance measures of representational space and behavioural data

Chapter main message

In the previous chapter, the effects of familiarity on representation in a DNN were explored by looking at the difference in Euclidean distance between image pairs for faces the DNN was either trained or untrained on. In this chapter, I look at other commonly used measures in computer vision: Cosine, Manhattan, and Jaccard distance. In this study, Jaccard distance is calculated in two ways, to mimic a generalized representation of a face (referred to as Jaccard-Global), versus a more detailed representation (referred to as Jaccard-Detail). The Cosine, Manhattan and Jaccard-Global distance showed the same pattern of results as the Euclidean distance, as discussed in Chapter 8 - Exploring the representation of familiar faces using deep neural networks., Experiment 7b. These findings are similar as the findings in Experiment 5a and 5b, indicating that the changes in representation after familiarisation are related to increased sensitivity to change in familiar faces. The pattern of results for the Jaccard-detail measure was almost as similar as the other measures, but lacked a significant interaction effect between Transformation Level and Familiarity for Transformation Level 5.

9.1. Introduction

In the previous chapter, the effects of familiarity in a DNN were explored by looking at the difference in similarity scores of image pairs between familiar (trained) and unfamiliar (untrained) faces. The similarity score used was Euclidean distance. As mentioned before, Euclidean distance is a measure that is often used in psychology to calculate similarity. However, there are other measures that are commonly used in machine learning: Cosine, Manhattan, and Jaccard distance. Each of these measures has a different calculation and different characteristics. To gain greater insight into human similarity and dissimilarity ratings and understand how these measures relate to each other, I will explore these different distance measures of representational space and their relationship with the behavioural data. Specifically, I compare how well these measures follow the pattern of the behavioural data found in Experiment 4 “Familiarity and similarity” and Experiment 5 “Familiarity and dissimilarity”. Understanding the different

calculations that best reflect the data will provide further insight in how we use the information we have stored in face space, but also what measures are useful for future research.

Euclidean distance

To recap, since the results of the measures will be directly compared to the findings of Experiment 7b (Chapter 8 - Exploring the representation of familiar faces using deep neural networks.) using Euclidean distance. The Euclidean distance is the distance between two points in Euclidean space, in which Euclidean space represents any (positive) N-dimensional space. This space is a vector space that consists of real numbers and values are continuous. The distance between two points is calculated using the Pythagorean Theorem, which states that the squared sum of the two shorter sides of a triangle (side a and b) is equal to the length of the longest side (the hypotenuse, c) squared ($a^2 + b^2 = c^2$). Any additional dimensions will be perpendicular to the existing planes. Thus, the Pythagorean Theorem in multidimensional space generalizes to $\sum_{i=1}^n d_i^2 = \text{hypotenuse}^2$. Where d is the vector length between point a and b in the *i*th dimension.

As briefly discussed in Chapter 1, the concept of ‘face space’ and the idea behind face space (Valentine, 1991) is that research on distinctiveness, inversion and race all indicate the influence of knowledge of the population of faces a person encountered so far. Thus, he proposed a framework that can encode a face as a point in a multidimensional space that is based on previously encountered faces. The dimensions in this space represent features that are used to encode a face. The space was proposed to be Euclidean in the example, indicating that the distance between points in this space relate to similarity of faces: the more similar a face, the closer it will be in face space. However, this was more to illustrate the principle as the author did mention that the “Euclidean metric will almost certainly be an over-simplification”. Therefore, I will explore other distance measures and see how well they reflect the pattern of behavioural findings.

Cosine distance

Overall, there has been little research done on relating image-based measures of similarity to perceived similarity. However, out of the three measures mentioned here, the Cosine similarity has been proposed multiple times. Valentine (1991) suggested the angle between vectors in face space is relevant, in which the vector lies between the origin of face space and the point that

represents the face. For similarity between two faces, calculating the angle between these two vectors will reflect the distance between the points of two faces. The distance from the norm is added to the model to account for effects of distinctiveness.

Whereas the Euclidean distance can become as big as it needs to be to reflect the difference, the cosine of an angle is limited to have a value between -1 and 1. Thus, whereas the Euclidean distance reflects the difference between face space dimensions and the values on these dimensions, the cosine distance reflects only the difference between face space dimensions. Comparing the pattern of results of Euclidean and Cosine distance will provide us an indication about the use of qualitative (differences between dimension) and quantitative (differences in values on the same dimension) differences in familiar and unfamiliar face recognition. As the angle between two vectors gets smaller and goes to zero, the value of the cosine similarity goes to 1. Cosine similarity can have a value of minus 1, when the angle between two vectors is 180 degrees. However, this is unlikely when the number of dimensions increases, as the values per dimensions will each have a normal distribution, causing the average to settle around 90 degrees. The lower the angle, the closer the vectors are to each other, thus the more similar they are. This limited range of the values allows for comparisons to other cosine similarities and setting a threshold.

Manhattan distance

The Manhattan distance is the sum of the absolute difference between 2 vectors. This is similar to the Euclidean distance except the Euclidean distance is the square root of the squared differences between two vectors. The Manhattan distance is also known as the L_1 norm and the Euclidean distance is also known as the L_2 norm. In a generalized way, the L_k norms allow for different ways to calculate distances based on vectors in a space. Thus, Manhattan and Euclidean distance are variations of such norms. Due to the mathematical resemblance of the Manhattan distance to Euclidean distance, it is likely that they both have a negative correlation with perceived similarity. Although the Manhattan and Euclidean distance are similar, Aggarwal, Hinneburg and Keim (2001) found that the Manhattan distance has better discriminative ability in high-dimensional spaces than the Euclidean distance. However, the effect of the k -parameter on discriminative ability is also dependent on the distribution of values per dimension. Thus, the

Manhattan distance is included to explore if the Manhattan has better discriminative abilities than the Euclidean distance with the distributions present in the current data set. As the Manhattan distance is a commonly used distance measure in computer vision, understanding if these distances result in a similar pattern is useful for comparing results across different studies.

Jaccard distance

To our knowledge, the relationship between human face recognition and Jaccard similarity have not yet been explored. Whereas most other measures discussed use the entire vector in a space as a point and calculate the distance between those points, the Jaccard similarity compares all components of the vector, that are non-zero in at least one of the vectors, and bases similarity on the total number of components that are equal. The formal definition of the Jaccard similarity is the intersection over union. Due to the different nature of the calculation, it is difficult to predict how this measure will relate to perceived similarity. However, any distance measure is likely to have a negative correlation with perceived similarity.

Jaccard similarity is used here to simulate the range of acceptable values on a dimension in a face representation, as per the conceptualization of Face-space-R by Lewis (2004). To recap my brief discussion of section 1.7 ‘Exploring face representations through (psychological) models’, face space is the “multidimensional space framework” by Valentine (1991) that describes an organizational space in which faces are represented as a location defined by its value on multiple dimensions. Face-space-R is a computational implementation of Valentine’s (1991) face space by Lewis (2004). In this model, activation of a face can be represented by a Gaussian radial basis function (RBF; Broomhead & Lowe, 1988) on each dimension. Each face has a certain optimum value for each dimension, and the activation reduces when you move farther from that optimum, and this follows a Gaussian distribution. The idea of face space has many unknowns, such as what the dimensions are, how many there are, etc. One of these unknowns is also the characteristics of these RBFs, such as the consistency of shapes over different dimensions for each face, or the scale of the variation between faces.

An additional unknown is the standard deviation of the RBFs, which has been held constant in the simulations performed by Lewis in his proposal of face-space-R (2004). As noted by Hancock (2021), face-space-R does not incorporate learning. Thus, it is unclear how the shape of the gaussian RBF changes when one becomes familiar with a face. Variability appears

to be integrated in the representation of familiar faces but not of unfamiliar faces, leading to familiar faces having a more robust representation than unfamiliar faces (Burton et al., 2005; Hancock et al., 2000; R. J. Johnston & Edmonds, 2009). One could argue that ‘robust’, in the context of incorporating variability, could mean a wider gaussian RBF in face space, i.e. familiar faces have a larger spread of dimensional activation compared to unfamiliar faces. Alternatively, it could be that a representation of a familiar face becomes more ‘robust’ by having a narrower RBF on dimensions that are critical for recognition, i.e. dimensions that are distinctive for the recognition of the identity. In this study, the Jaccard similarity is used to mimic changes in RBFs in order to explore task differences in the behavioural data.

As mentioned above, to calculate the Jaccard similarity you take the total number of components that are equal between both vectors, ignoring any that are zero in both, and divide by the total number of attributes/values you are comparing, i.e. the vector length. The activation value of a feature in a representational vector is a continuous number, and no matter how small the difference, it will count it as dissimilar if it is not exactly the same value. Thus, in a practical sense, to get any degree of similarity between two long vectors with continuous values, it requires rounding of the values in the representational vector, i.e. a threshold to decide which values should be considered as equal. In order to mimic changes in RBFs through Jaccard similarity, different rounding of the values in the representational vector are used to represent different levels of precision. For an example of Jaccard distance, please see Figure 51. In the first example, two-decimal values of two vectors are compared. Although similar values for the majority of the vector, they are not exactly the same, thus the Jaccard similarity value is 0.2 (one pair is the same). In the second example, all values of example 1 are rounded to 1 decimal, and the similarity becomes 0.6.

Example 1: 2 decimals numbers					
vector 1	1.12	1.32	1.67	1.74	1.50
vector 2	1.13	1.24	1.81	1.66	1.50
equal?	0	0	0	0	1
similarity:	$(0+0+0+0+1)/5=$				
distance=	1-similarity				

Example 2: rounded numbers					
vector 1	1.1	1.3	1.7	1.7	1.5
vector 2	1.1	1.2	1.8	1.7	1.5
equal?	1	0	0	1	1
similarity:	$(1+0+0+1+1)/5=$				
distance=	1- similarity				

Figure 51 Example of the rounding for the Jaccard computation.

Using different precision to round down the values in the representational vector, I aim to gain deeper insight into the two main findings presented in this thesis: familiar faces are perceived as more similar (Experiment 4a), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). As discussed extensively in this thesis so far, it is likely that the difference between familiar and unfamiliar faces lies within the representation of the faces in face space.

In terms of the Jaccard similarity, a focus on similarity over dissimilarity could be represented as a global comparison of two face vectors, in which global is operationalized as a comparison of values that are equal when the accuracy of the description is decreased (more details in the Methods section). In the case of focussing on dissimilarity, one would still want to have the values rounded, as we never expect a face to be exactly the same due to viewing conditions, but one would like to retain as much information as possible to find the differences. In the terms of Face Space-R, two vectors being the same can be interpreted as both being within the radius of the radial basis function for that dimension. Rounding to fewer digits corresponds to having a wider radial basis function. In this study, I have chosen two different rounding factors to reflect both these situations, in which the lowest number of decimal places represents a global comparison focussed on similarity and the highest number of decimal places represents a detailed comparison.

9.2. Methods

The methods of this experiment are the same as Experiment 7b in Chapter 8 - Exploring the representation of familiar faces using deep neural networks.’ and the distant measures are calculated on the same data used. A brief description of the calculation of each distance measure is given in the following section, discussing the unique characteristics they have in comparison to each other. For the computations of the similarity ratings, the ‘`scipy.spatial.distance`’ function is used from the SciPy library (Virtanen et al., 2020). Therefore, all similarity ratings are in their distance form, meaning a higher value reflects a higher degree of dissimilarity.

Cosine distance

To calculate the cosine similarity, two vectors are projected in a multi-dimensional space and the cosine of the angle between the two vectors reflects the similarity. The software library I use, SciPy (Virtanen et al., 2020), calculates the cosine distance by taking 1-cosine similarity. Similar to concepts about face space (Lewis, 2004; Valentine, 1991) a vector in the image space has a magnitude and a direction. In our case, there will be 2622 dimensions, corresponding to the 2622 features in the representational vector produced with VGGFace (as there were 2622 identities present in the first VGG image set).

Manhattan distance (and Minkowski distance)

The Manhattan distance is the sum of the absolute difference between 2 vectors. This is similar to the Euclidean distance except the Euclidean distance is the square root of the squared differences between two vectors.

Another common measure in computer vision is the Minkowski distance. The Minkowski distance can be the same as the Euclidean and Manhattan distance, as it takes the sum of the absolute difference between 2 vectors too but has the addition of a p-value that is used to raise the difference between 2 vector points by the power of p, and the summation of those values is then raised to the power of 1/p: $D(X, Y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$. The Euclidean distance is also referred to as L2 norms and the Manhattan distance is referred to as L1 norms, referring to the value of p. Thus, if p equals 2, the Minkowski distance is equal to the Euclidean distance, and if p equals 1, it is equal to the Manhattan distance. However, any value over 1 is a valid value for

p, but a value of 1 or 2 is most commonly used. For this reason, the Minkowski distance is not included in this chapter.

Jaccard distance

The Jaccard distance is calculated by dividing the intersection of the two vectors by the total attributes compared, i.e. the vector length. This means the number of all the values that are the same between vectors are the intersection and these are divided by the total number of values in each vector, ignoring any values that are zero in both vectors. The higher the number, the greater the similarity. In the SciPy library (Virtanen et al., 2020), the distance is used, which is equivalent to 1 minus the similarity.

In the application to faces, the likelihood of components being exactly the same is extremely low. As mentioned in the introduction of this chapter, I have chosen two different rounding factors to reflect both these situations, in which the lowest number of decimal places represents a global comparison focussed on similarity and the highest number of decimal places represents a detailed comparison. These will be referred to as Jaccard-Global and Jaccard-Detail respectively. The values given in the example, 0.0367435463 and 0.0367435462, are of the magnitude present in the vector generated by the VGGFace model. Thus, the Jaccard-Global will be the Jaccard distance between vectors that have their values rounded to 3 decimals, and Jaccard-Detail will be the Jaccard distance between vectors that have their values rounded to 4 decimals. For the given example, the value in the Jaccard-Global vector will be 0.037 and the value in the Jaccard-Detail will be 0.0367.

Analysis

To assess whether these distance measurements result in the same pattern for the DNN vector comparison as the human data, the same LMM as in Chapter 8 - Exploring the representation of familiar faces using deep neural networks. is used. To recap, the model included Familiarity and Transformation Level as categorical factors (Familiarity: familiar and unfamiliar coded as -0.5 and 0.5 respectively to allow for the analysis of the interaction; Transformation Level: 2, 3, 5, or 7), and the intercept, slope and interaction of Familiarity and Transformation Level in the performance per model as a random effect. Due to convergence issues, the main effects were excluded from the random effects in the case of the Cosine and Manhattan distance. The

parameter estimate (b), standard error (SE), t value and p value are reported. The linear mixed effect analysis was performed in R (R Core Team, 2019) with *lmerTest* (Kuznetsova et al., 2017). The threshold for significance was set at $\alpha = 0.05$. Normality of the residuals on the model were inspected visually and showed no violation.

9.3. Results

Before I discuss the results, I will recap the three findings that are important to the interpretation of the results. These three findings are the results of Experiment 4a and 4b (Chapter 0 and 4), Experiment 5a and 5b (Chapter 5), and Experiment 7b in Chapter 8. As mentioned previously, the findings in Experiment 4 - familiar faces are perceived as more similar (and not interaction effect) - are possibly explained by the idea that people focus on finding similarity instead of focussing on dissimilarity. Whereas the findings of Experiment 5 - the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (in addition to a main effect of Familiarity) – the emphasis should be on dissimilarity. These findings are in line with the finding that ‘telling faces apart’ and ‘telling faces together’ seem to be dissociated skills (Megreya & Burton, 2006)

In Experiment 4a and 4b, the effect of familiarity on the perception of multidimensional transformations was explored. The stimuli consisted of these 8 average images and variations that had either 2, 3, 5 or 7 colour components (components 10 and 11; 10 – 12; 10 – 14; 10 – 16 respectively, referred to as ‘Transformation Level’) in the PCA space changed. Participants performed a face matching task in which in each trial, an original face image was shown, followed by a mask, the altered image and another mask. Finally, the participants were presented with the question “Could these be two pictures of the same person? 1 – certain no; 2 – think no; 3 – guess no; 4 – guess yes; 5 – think yes; 6 – certain yes”. A main effect of both Transformation Level and Familiarity was found, but there was no significant interaction effect.

In Experiment 5a and 5b, I used PCA transformations explore the effects of familiarity on sensitivity to change. A sequential 2-alternative forced-choice task was used to explore differences between familiar and unfamiliar faces in the ability to identify changes made to the appearance of the face. The task was carried out with transformation of 2SD (Experiment 5a) and 4 SD (Experiment 5b) changes. In each trial, the original was shown, followed by a mask, the

first altered image, another mask, the second altered image and a last mask. Finally, the participants were presented with the question “Which of the two changed pictures more closely resembles the person shown in the first image 1 - first picture; 2 - second picture”. All transformation levels of the previous experiment were used, and all possible combinations between image pairs were used for all participants (an image with 2 components changed paired with an image with 3 components changed, 2-5, 2-7, 3-5, 3-7 and 5-7). The effect of Pair Distance and Familiarity were explored, in which Pair Distance was the difference between the number of components altered in the two images (so for example, if in a trial an altered image with 2 components changed is presented, and another altered image with 5 components, the Pair Distance is 3). There was a significant main effect for Pair Distance and Familiarity, and a significant interaction effect between Pair Distance and Familiarity, but only for the largest Pair Distance.

In Experiment 7b, a method of familiarization of faces in the form of fine-tuning of a pre-trained DNN is used to explore the two main findings mentioned above: familiar faces are perceived as more similar overall than unfamiliar faces (Experiment 4a), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). To get similarity ratings from the DNN that can be compared to the human similarity ratings, all images are fed through the model in order to generate an image-vector from the representational layer. Then for all pairs (a pair of an original image and one image with either 2, 3, 5 or 7 components changed) the difference between the vectors is calculated using Euclidean distance. A significant main effect for Transformation Level, but not for Familiarity, was found. Additionally, there was a significant interaction between the two. These findings are similar to the findings in Experiment 5a and 5b, indicating that the changes in representation after familiarisation are related to increased sensitivity to change in familiar faces. The descriptives of these experiments are shown in Figure 52 (Experiment 4a, 5a and 5b), and in Figure 53 (Experiment 7b).

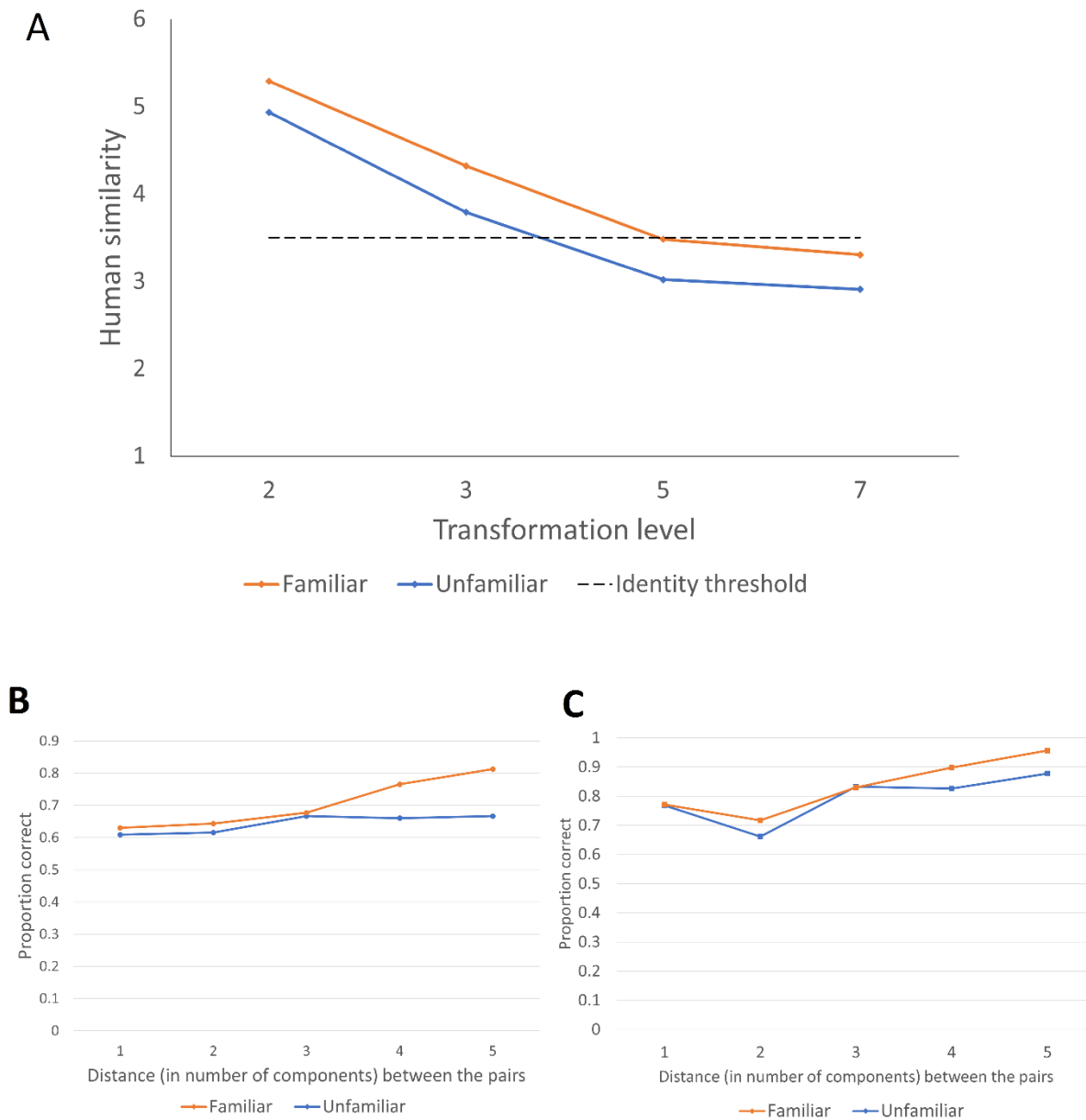


Figure 52 Recap of the findings of Experiment 4a, 5a and 5b.

Manhattan, Jaccard-Global and Jaccard-Detail distance all showed a similar pattern of results as found for the Euclidean distance in Experiment 7b: a significant main effect for Transformation Level but not for Familiarity, and a significant interaction effect between Transformation Level and Familiarity for Transformation Level 5 and 7. The only exception was for the Jaccard-Global and Jaccard-Detail distance: the interaction effect between Transformation Level and Familiarity

was already significant at Transformation Level 5. For an overview of all the comparisons and related p-values, please see Table 18 in the Supplementary materials – Chapter 8. These findings are similar as the findings in Experiment 5a and 5b, indicating that the changes in representation after familiarisation are related to increased sensitivity to change in familiar faces. Cosine similarity did not show a similar pattern of results as found for the Euclidean distance in Experiment 7b: There was only a significant main effect for Transformation Level but not for Familiarity, and no significant interaction effect between Transformation Level and Familiarity for either Transformation Level 5 and 7.

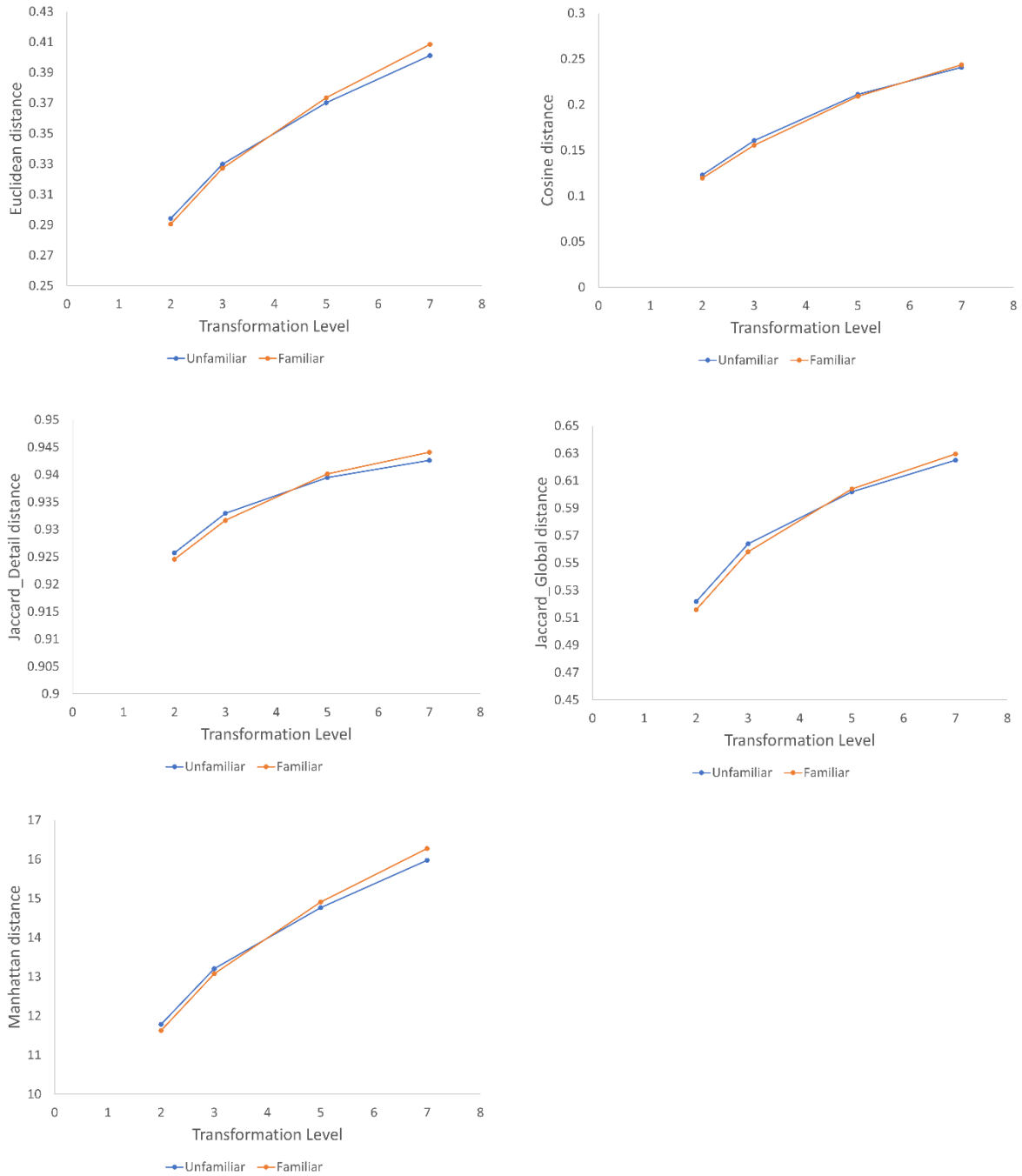


*Figure 53. Plots of the different distance measures per Transformation Level.
Please note that the y-axis of the Jaccard_Detail graph does not start at zero, for visualization purposes (as the interaction effects are significant).*

9.4. Discussion

In this chapter, the relationship between three different distance measures in representational space and behavioural data was explored. Specifically, I compare how well Cosine, Manhattan, and Jaccard distance follow the pattern of the behavioural data found in Experiment 4a (Familiarity and similarity) and Experiment 5 (Familiarity and dissimilarity) and how they compare to the Euclidean distance. From these measures, the Manhattan, Jaccard-Global and Jaccard-Detail distance showed the same pattern of results as the Euclidean distance, as discussed in Chapter 8, Experiment 7b. These findings are similar as the findings in Experiment 5a and 5b, indicating that the changes in representation after familiarisation are related to increased sensitivity to change in familiar faces.

The finding that the Manhattan distance showed the same pattern of results as the Euclidean distance should come as no surprise, as they are mathematically close. This measure was included to confirm this assumption, as this measure is amongst the most commonly used in the field. Similarly, the finding that Cosine distance does not follow the same pattern of results as the Euclidean distance is expected, as the Euclidean distance can become as big as it needs to be to reflect the difference, the cosine of an angle is limited to have a value between -1 and 1. However, the Cosine distance is commonly used in the field of computing science for image comparison. Thus, the finding that it does not reflect the same pattern as other measures and behavioural responses is important knowledge. The Cosine distance only showed a main effect for Transformation Level, which is ideal from a technical perspective, as it performs as well on untrained (unfamiliar) images as on trained (familiar) images. It appears that the differences in representation with regards to familiarity are quantitative difference and not qualitative ones.

The finding that Jaccard-Global and Jaccard-Detail shows a similar pattern as the Euclidean distance is unexpected since the nature of the calculation for the Jaccard distance is quite different from the Euclidean distance. Their similar result might be related to the obvious practical limitation in the exploration of the Jaccard distance: the adaptation of the rounding factors merely used two different rounding factors. This rounding factors were chosen based on the features values and to gain a general idea of the effect of different rounding factors, that in this case differed by a factor of 10. Although the pattern of the data visually changed, the pattern of the statistical results remained similar. Different rounding factors might better reflect the idea of differences in RBFs.

Additionally, the rounding was done for all dimensions, and the pattern might reflect the behavioural data better if the rounding was only applied on the distinctive dimensions. However, since these are still unknown, it should be clear that this is no claim of a full mechanistic explanation, but simply a mathematical translation of the idea of differences in RBFs. It should be obvious that the same can be achieved in many different ways, for example, by adding and altering threshold values. However, conceptually, this study shows that that variation in the error margin could explain behavioural data and that this idea can be implemented without the need of changes in face space/representational vectors.

At the moment, the findings of this chapter and Chapter 8 (Experiment 7a – c) are comparing behavioural data to one specific task and one particular set of transformations. To further validate these findings, further research is needed using other task designs, facial transformations, identities, architectures, etc. Additionally, I have only looked at two different architectures and the vectors used in this chapter was long and had very low feature values. Without further research, it remains unclear how the length of such vectors and other architecture-related aspects influence the results of experiments following this paradigm. Although it is useful to see that the pattern of results is similar, to further interpret the found relationship between image-based and perceived similarity, the biological plausibility of formulas and networks should be more integrated. I will further discuss this in the general discussion (Chapter 10).

Understanding the different calculations that best reflect the data provides a possible suggestion of how we use the information we have stored in face space, but also explains how different tasks lead to different findings and can still have the same underlying change. A deeper understanding of the measures that best reflect the task performance can also be applied in task design. It is not uncommon for a DNN similarity score to be used in estimating task difficulty. The findings related to Jaccard similarity show how we can alter these estimations in a fashion that is more tailored to the task.

10. Chapter 10 - Conclusions and future directions

In this thesis, I explored what internal representations of faces we have and how these are different between familiar and unfamiliar faces. From the review in Chapter 1, it should be clear that many open questions remain about the internal representation we hold of faces. In the theoretical models of human face recognition (Bruce & Young, 1986; Burton et al., 1990, 1999; Lewis, 2004; Valentine, 1991) the one factor explicitly identified to be missing is what information is stored in the needed representation for face processing. A better understanding of this internal representation will allow for the development of tasks that measure face processing abilities in more detail and increase our understanding of the visual information needed to identify a face, either familiar or unfamiliar. I have used two approaches: behavioural and computational. In this final chapter, I will summarize my main findings and identify several future directions that could build on the framework provided within this thesis.

10.1. Summary of chapters

A brief overview of the chapters, experiments, and findings can be found in Table 29 in Supplementary materials – General discussion. In Chapter 2, I examined a new paradigm in which participants could alter certain features in a gradual scale, to explore accuracy of an internal representation. In Experiment 1, I checked if the percept between familiar and unfamiliar faces was the same by showing the original on-screen and asking participants to match the image to the original. In Experiment 2, I repeated Experiment 1, but added a delay between the original and the response image, so that face memory needed to be utilized for the task. In Experiment 1 and 2, the features that were changed were distance between the eyes and the eyebrows, width of the mouth, lip thickness and brow thickness. The findings indicate that perceptual sensitivity but not familiarity has an influence on the accuracy of our face perception, but neither influence the accuracy of recall of the selected manipulations.

In Chapter 0, I presented Experiments 3 and 4, that explored the relationship between perceived similarity and image-based spaces. Little research has been done on the relationship between perceived similarity and image-based spaces which makes it difficult to quantify face transformation, leading often to the use of verbalizable transformations. By comparing the effect of single- and multi-dimensional changes on perceived similarity and 3 different image-based spaces, these experiments were used to assess the usefulness of PCA transformations in the

exploration of internal representations of the face. In Experiment 3, we found that perceptual similarity and the three image-based spaces are linearly related. In Experiment 4a, our findings show that there is a positive relationship between the number of components transformed and both the perceptual similarity and the same three image-based spaces used in experiment 1. Furthermore, we found that familiar faces are rated more similar overall than unfamiliar faces.

In Chapter 4, I explored if the pattern found in Experiment 4a would be the same when the changes were smaller (Experiment 4b; 2 SD instead of 4 SD) in terms of magnitude, to better understand the relationship between the number of components and magnitude of change, and to explore if the pattern we found in experiment 4a would be the same. Our findings show a similar result as in experiment 4a, but, with higher overall similarity scores while the difference between changes with 5 and 7 components is significant.

In Chapter 5, I explored the relation between familiarity and sensitivity to change. Two experiments are presented, using a two alternative forced choice task to explore differences between familiar and unfamiliar faces in the ability to identify changes made to the appearance of the face. The task was carried out with transformation of 2SD (Experiment 5a) and 4 SD (Experiment 5b) changes. The different magnitudes of changes were used to explore effects of cross-identity versus within-identity changes. A main effect of perceptibility of change (in number of components) between the response pairs was found. Additionally, I found a significant main effect of Familiarity and an interaction between familiarity and perceptibility of change. This means that the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces. Both experiments revealed a similar pattern of results.

Chapter 6 explores the boundaries of identities further, by exploring the relationship between familiarity and the consistency in perceived change. In this Experiment (6), different sets of components were changed and presented to participants in a 1-to-4 array. Participants were then asked which of the four changed pictures they thought had changed the appearance of the person the most. The results of Experiment 6a in relation to familiarity were difficult to interpret due to a large majority being familiar and unfamiliar with the same identities. There appeared to be a preference for a certain component dependent on the identity, one that cannot be reasoned away by the size of the change measured in an image-based space (Gabor distance). The results of Experiment 6b show that component 12 has the strongest effect on appearance in all identities, though the preference is minor compared to the effects of components seen in

Experiment 6a. Since the same component affected all identities in a similar fashion, there were no clear indicators that familiarity influenced the distinctiveness of transformation.

In the literature review of part II (Chapter 7), I discussed the work that has been done using deep convolutional neural networks to explore human face recognition as background to Chapter 8, that presented the computational experiment in which I mimic familiarity by training on certain identities and comparing the representational spaces of familiar (trained) and unfamiliar (untrained) faces. A method of familiarization of faces in the form of fine-tuning of a pre-trained DNN is used to explore the two main findings presented in the thesis: familiar faces are perceived as more similar overall than unfamiliar faces (Experiment 4a and 4b), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). Three experiments were run: in the first experiment the FaceNet architecture was used with 341 layers frozen and 106 layers trained (Experiment 7a) and without any frozen layers to achieve maximum performance on the test set (Experiment 7c), and the VGGFace architecture was used with 30 layers frozen and 8 layers trained (Experiment 7b). In Experiment 7a, no effects of familiarity were found. In Experiment 7b, a main effect for both Transformation Level and Familiarity was found. Additionally, there was a significant interaction between the two. Lastly, in Experiment 7c, similar to Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found, there was only a main effect of Transformation Level.

In Chapter 9, I further explored the results of Experiment 7b and compared the representational spaces using different distance measures, to find which outcome measure best reflects a human behavioural pattern. Cosine, Manhattan and Jaccard-Global distance all showed the same pattern of results as found for the Euclidean distance in Experiment 7b: a significant main effect for Transformation Level but not for Familiarity, and a significant interaction effect between Transformation Level and Familiarity for Transformation Level 5 and 7.

10.2. Difference between familiar and unfamiliar faces

The overarching aim of this thesis was to explore what internal representations of faces we have and how these are different between familiar and unfamiliar faces. The work presented in the previous chapters collectively show four main findings: there is a linear relationship between image-based spaces and perceived similarity that can be used to explore the effect of PCA

transformations, familiar faces are perceived as more similar overall compared to unfamiliar faces, the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces, and that the effect of familiarity can be found at the representational level (in DNNs).

The first main finding mentioned – there is a linear relationship between image-based spaces and perceived similarity – should flag the fundamental knowledge that is still missing on the role of image-based properties and its role in the field of face recognition. As I will discuss in the section ‘Future directions’, the approach of computational cognition will become increasingly important, and a better technical understanding of visual stimuli (understanding the role of image properties such as dynamic range, contrast energy, colour intensities, etc) will be necessary to benefit from interdisciplinary approaches. The remaining three major findings can be considered a testament to the success of this approach.

10.3. Future directions in computational face recognition research

In the next section, I will discuss the most promising future directions and how the work presented in this thesis could contribute.

Understanding face learning

One clear aspect that needs further exploration is face learning: how do we go from encountering a face for the first time to a robust representation of a highly familiar face? In this thesis, little attention has been given to the process of building a face representation. Instead, this issue was simply avoided in the behavioural experiments by using familiar versus unfamiliar faces, as is a common strategy in face recognition research. Although this is the go-to strategy, it is a limitation. The importance of understanding is clear and there is work that attempts to understand variety through broader categories (of high versus low variability for example). More insight in the learning process will provide new insights in the information that is stored in the representations of faces. Although addressed to a limited extent in this thesis, learning is intimately related to exploring internal representations, thus, I'll now consider how each of the four principal findings relates to learning.

Firstly, the finding that there is a linear relationship between image-based spaces and perceived similarity can be used to address a current limitation in how variability is defined: currently, the most common approach is to compare low and high variability, due to the complex nature of face images. The found relationship between image-based spaces allows for a finer definition of variability. In an alternative approach, the number of images is increased during face study, to increase the amount of variability the participant is exposed to of certain faces. In these types of paradigms, the relationship between image-based spaces and perceived similarity can be used to quantify the increase in variability, which would make it easier to compare different identities and findings in other studies. In the current paradigm, I avoided the earlier components and chose later ones, that affect identity. However, different selections can be made, that affect identity to a different extent. For example, if earlier components are chosen, which often capture lighting or angle, this could provide for variability that does not affect identity. Creating different types of variability in this paradigm allows for a deeper exploration of the types of variability useful for learning.

Secondly, the two behavioural findings - that familiar faces are perceived as more similar overall compared to unfamiliar faces and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces – can lead to a greater understanding of the mechanisms of increasing familiarity over time. In this thesis, the effect of relatively subtle face changes was tested, thus these findings could most likely be replicated with a larger effect size if bigger transformations are presented. A larger effect size would make it easier to track a changing representation. However, in order for this to work, it is likely that the most distinctive features need to be defined. For this, the fourth major finding presented in this thesis can be used.

As concluded in Chapter 0, the difficulty of the PCA framework as described in this thesis is that the transformations are specific to the PCA space calculated, and that different PCA spaces do not align to one another. This makes the framework a poor tool to identify general distinctive features. One would need a large PCA space based on a large number of identities, so that any new identities can be well represented within the existing PCA space, which is not feasible for several reasons. To overcome the limitation of generalizing to other identities and ensuring all identities can be represented as well as possible, Deep Neural Networks (DNNs) can be used. As should be evident for the literature and findings presented in Chapter 8, there currently is no clear paradigm to explore familiarity in DNNs. An additional challenge is that

there appears to be a relationship between distinctiveness and familiarity, which need to be better understood if one wants to understand the effect of familiarity on internal representation.

The methods behind the fourth major finding - and that the effect of familiarity can be found at the representational level (in DNNs) – provides a paradigm to explore what hyperparameters (parameters related to training a DNN) can be used to mimic familiarization and gain a better understanding of human face perception. The strength of this paradigm is that it builds on the two behavioural findings, as these provide a way to validate the found patterns in the DNN data. For the purpose of this thesis, only a couple of DNNs were explored. However, in order to explore familiarity and distinctiveness, further research is needed. Specifically, there are many other parameters in DNNs that are yet unexplored but likely would help to mimic human behaviour better. Additionally, only a handful of networks has been used in this line of research, and it is unclear what makes an architecture suitable for these paradigms. I will discuss each of these points in the following sections.

Exploring a wider range of machine learning techniques

One major contribution to the progress of DNNs was the improvement of the learning algorithm (Chapter 7). Although successful, there is of course a still-ongoing pursuit to make learning more efficient. As mentioned above, it should be evident by this point in the thesis that a lot remains unclear about face recognition, with an especially large gap in literature when it comes to how we learn faces. The emphasis in face processing, on the technical and experimental side, has been on the convolutional type of networks. Exploring a wider range of machine learning techniques will provide further insight in human face processing. There are many aspects of machine learning that can be explored, I will now discuss the recent developments that best fit the work presented in this thesis (including the computational approaches presented in the literature review in Chapter 7).

Other aspects of face recognition

I only explored the stage of representation in this thesis, and do not include the other aspects of face recognition, such as social face perception, identification and expression recognition. However, from the Bruce and Young model on human face recognition (Bruce & Young, 1986)

it is clear that the different aspects of face processing influence each other. Similar to the future research in face learning, it will become more important to understand where similarities exist, and where human and artificial face recognition deviate for each of the other aspects of face recognition involved. Although not directly related, the methodology in this thesis can inform experimental designs that are suitable for testing both psychological and computational research questions that are aimed at exploring the other aspects in face recognition.

Getting rid of the black box mentality: visualisation of activation

One of the major comments to the computational approach, is the black-box nature of DNNs: understanding what is going on inside the network and how training protocol and hyperparameters affect that is one of the major challenges. The assumption in a computational approach should not be that one model is built that can provide a conclusive answer, but that each build model informs the question that is asked and can be adapted to further define mechanisms (Box, 1979). However, although it presently does not matter that the DNNs are likely not human-like in their mechanisms, it is important to know how a model is deviating from human processing (Box, 1979; O'Toole & Castillo, 2021; see Chapter 7). The latter point, knowing how a model deviates, is one of the challenges faced in computational frameworks. Thinking back to handful of studies using DNNs to explore human face recognition and the experiments presented in this thesis, one might note it is mainly focused on the output of DNNs. The activation throughout the network is largely ignored, and the main focus lies on performance. The most recent development is the approach used in this thesis and by Blauch et al. (2021b), in which the representational vector (the very last layer before the classification layer) is used.

The reason for this current lack of exploration of the activation throughout the neural networks is technical in nature, but will most likely change soon. AlexNet was a pivotal moment and the complexity of available models starkly increased, and it has only been less than a decade since DNNs were feasible to use in academic research. After the first amazement about the performance of these networks has passed and additional small improvements were made, researchers are becoming more interested in how these networks are doing what they are doing. This has led to the development of techniques that visualize the contents of neural networks, such as gradient ascent (e.g. Erhan et al., 2009), which generates a synthetic image that depicts

what images would activate it the most. Probably the most famous example of the implementation of gradient ascent is Google's DeepDream (Mordvintsev et al., 2015). There are also techniques that not only show the generic activation of the node, but the node's activation to a specific input images, such as done by Samek et al. (2017) to evaluate what a neural network has learned.

Such visualisation of nodes could most likely be used to gain a better understanding of distinctive facial features. Once these are identified this can be combined with the first major finding in this thesis - that there is a linear relationship between image-based spaces and perceived similarity that describes PCA transformations well - to explore what mediates a change in perceived identity. However, it remains unclear to what extent a DNN exhibits human-like behaviour. For example, the distinctive features the neural networks might select, might not be relevant for humans. Recently, the interest in defining human-like networks has increased, which I will discuss in the section 'Brain-likeness paradigms and biological plausibility'.

Understanding input of training

Apart from more efficient learning algorithms, the progress in DNNs was also heavily influenced by the availability of data sets (Jia Deng et al., 2009). Although the quality of the data sets containing faces have vastly improved in the last decade, there is still a strong need for understanding the variety if we want to gain a better understanding of all the information we retain on faces. Currently, the handful of studies using a computational approach, use pre-trained networks that have been trained on millions of images. This makes sense, as the goal from a computer vision perspective is to get the best performing networks. If one wants to use these networks, the input images need to be considered as a parameter one can tweak (similar to hyperparameters in model training) that can mimic more human-like responses. Blauch et al. (2021b) provides a framework for this, as they varied the exposure to facial images to simulate to 'experience' with a face. However, they note that they never attempted to make a human-like responding network, thus the findings should be interpreted as such.

Another direction of research is using the relationship between image-based spaces and perceived similarity to define other suitable approaches to create systematic facial transformations. The experiments in this thesis and most of the research in face recognition uses 2D-images and PCA representations, thus the findings should just be interpreted for 2D-images

and PCA but it remains unclear to what extent these findings extend to, for example, transformations that are created with a different technique than PCA, such as a GAN. Other types of transformation could lead to new insight in face space.

One could argue, that to truly understand face representation we must understand the entire representation of a person and all interaction between available information (i.e. all components in Bruce and Young's model, 1986). Although a discussion on modality in the brain would go far beyond the scope of this thesis, it is important to point out that some do believe that the approach of complete-brain modelling is the next step forward. Models of the entire brain are attempted (e.g. the Blue Brain project, <https://bluebrain.epfl.ch/>), and it will be interesting to see what insights we can gain from larger networks.

Again, the caveat is that for a better understanding of training and input, it would be preferable to have a more human-like responding network. Understanding what parameters, and at this moment in time also what architectures, are best used to mimic human face recognition is far from known. Finding these parameters needs a systematic review of comparing different architectures and parameters, which is currently very time-consuming. However, in the (somewhat) near future, I expect this to change; just like the jump in processing power there was going from CPU to GPU, the move from classic to quantum computer will bring an equivalent step forward in modelling face recognition.¹ Although time-consuming, some research has already started such reviews, focussing on 'brain-likeness' and biological plausibility in neural networks (discussed below).

¹ To clarify, I don't expect quantum computers to be available for consumer soon, but I imagine more a google colab type of situation, or at least, I am hoping for one!

Brain-likeness paradigms & biological plausibility

How might the match of DNN to human performance be better assessed and further improved? Two related approaches in computational neuroscience exist, one focusses on assessing ‘brain-likeness’ and the other on biological plausibility.

Integrative benchmarking in computational neuroscience

In the context of object recognition, Schrimpf and colleagues (Schrimpf, Kubilius, Hong, et al., 2020; Schrimpf, Kubilius, Lee, et al., 2020) set out to create a platform to collect and assess computational models of the ventral stream and provide a ‘brain-score’. This brain-score is a score that captures multiple benchmarks, both neural and behavioural. At the time of publishing, the neural benchmark is about the resemblance between feature activations in an artificial neural network (ANN) and the neural activation in primates. However, many benchmarks have been added since, with at the time of writing multiple benchmarks for V1, V2, V4, IT, behaviour, and engineering. The latter, engineering, is an umbrella term and refers to non-brain measures that can be related to brain measures, such as robustness to image perturbations as an indication for V1-likeness. They refer to ANNs, because the platform allows for more types of networks than DNN (which is short for deep convolutional neural networks), such as transformer networks. In the behavioural benchmark, outputs of the ANNs – such as predicted classes in an object categorization task - are compared to that of primates.

The idea behind the brain-score and the associated platform (brain-score.org) is the belief that the divide-and-conquer-later approach to understand visual processing in the ventral streams has proven limited in the ability to understand human intelligence (Schrimpf, Kubilius, Lee, et al., 2020). They argue that, to advance neurally mechanistic models of human intelligence, integrative benchmarking is necessary. The integrative part refers to the different benchmarks that go into the brain-score, these are made up of different datasets that were tailored for exploring specific subprocesses and/or areas. By combining these measures, new insights can be gained if models either work well on the other metrics, or fail. The authors (Schrimpf, Kubilius, Lee, et al., 2020) do mention explicitly that they do not believe that they have the perfect metrics to calculate the current version of the brain-score, but the idea is that this will change with the addition of different metrics (that other labs can submit).

Although an exciting concept and platform, the usefulness for understanding face perception specifically might be limited for the time being, as many of the models have focused on lower-level areas from the ventral stream. Additionally, the behavioural benchmark only includes object categorization tasks at the moment, although the platform does allow other benchmarks to be submitted, and the owners of the site have expressed interest in the sort of face similarity tests developed here (Schrimpf, personal communication). The concept of an expanding platform is in line with the idea of informative iteration by Box (1979; Chapter 7): “scientific progress comes from motivated iteration between theory and practice.” In the following years, metrics for face processing that can be used in the behavioural benchmark need to be developed, compared to the other metrics, and this comparison should be used to inform new experimental design. The work presented in Chapter 8 - Exploring the representation of familiar faces using deep neural networks.’, could be used as a basis for such a metric.

Brain hierarchy

Although the brain-score is a composite score of all the different benchmarks that are based on different metrics, the relationship between the subprocesses and areas and the change of the representation is not included in the score (Nonaka et al., 2021). The layers in a neural network represent increasingly complex visual structure in a similar fashion to the human system (Chapter 7). Nonaka et al. (2021) argue that, due to the importance of multistage hierarchical processing in perception and cognition, assessing the hierarchical homology between brain and DNN in a quantitative manner could be a way of estimating functional similarity. They translated this into the brain hierarchy (BH) score, in which the alignment between DNN and brain hierarchy is compared through encoding and decoding approaches.

Decoding refers to the process of training a network to find the relationship between the brain activity and certain visual stimuli. In other words, a network is trained to predict what stimulus was presented based on the brain activity. For the purpose of the BH score, Nonaka et al. (2021) used a decoding approach proposed by Horikawa and Kamitani (2017) that uses hierarchical visual features to decode. They defined their regions of interest as V1, V2, V3, V4 and subsequent areas are included under the category referred to as ‘higher visual cortex’. For each of the DNNs, fMRI data is decoded and for DNN units in each layer (a thousand units were selected if the layer contained more units), the top region of interest is decided based on the

highest correlation value. In addition to the decoding measure, encoding was used. Encoding is similar, but now the DNN unit activations are used to predict voxel activation. The two scores together make up the BH score.

Nonaka et al. (2021) then proceed to use the BH score to explore if there are architectural characteristics associated with the degree of hierarchical homology between the brain and 29 DNNs. As there are many architectural characteristics in modelling, they focus on five components: the presence of fully-connected layers, branch connections and skip-connections, and the number of layers and weight parameters. Their paradigm allows for the exploration of the effects of these architectural characteristics on hierarchical homology. This is useful for the selection of an architecture when exploring human perception.

Again, however, the usefulness for understanding face perception specifically might be limited for the time being, as this technique uses the hierarchy of the areas involved as an indication for brain-likeness and any areas beyond V4 are only included under the category referred to as ‘higher visual cortex’. This means there are no hierarchical levels to distinguish for any subprocesses of face recognition that is not performed by V1, V2, V3 or V4. A better understanding of the visual stimuli used in face recognition tasks and the activation in the areas beyond V4 could help develop tasks more suitable for the exploration of hierarchical homology between the brain and DNNs. The finding that there is a linear relationship between image-based distances and perceived similarity can inform experimental designs aimed to explore such topics.

Biological plausibility

In addition to assessing the brain-likeness of artificial networks, there is an increasing interest in creating models that are more biologically plausible. As technology progresses, the architectural characteristics of DNNs that are not biologically plausible can be replaced by other characteristics that are more plausible. Pogodin et al. (2021) set out to address two mechanisms in DNNs that are biologically not plausible, but are often included in models of visual processing: weight sharing and convolutional connections. As a more biologically plausible alternative to weight sharing, they explore data augmentation and lateral connectivity in a locally connected network.

Data augmentation was used to mimic the multiple views one gets when making saccades. The problem with this approach is that it requires a high number of translations to improve the performance of a locally connected network compared to a convolutional network.

And even in this scenario, the improvement in the performance remained limited. Thus, the authors conclude that the number of translations is so large, that this does not correspond with the number of translations gained from saccades.

The second alternative Pogodin et al. (2021) explored was weight sharing through lateral connections. In a locally connected network, weights will diverge. To avoid this divergence and to make them convolutional, the training was interrupted to equalize the weights. This interruption was called the ‘sleep’ phase. The averaging was done by connecting every third neuron in input and output and applying a specific learning rule, which results in a convolutional network. In contrast to the data augmentation, this method led to large improvements in the performance of the network, and did not need more data but just the introduction of the ‘sleep’-phase. Pogodin et al. (2021) note that the results indicate that DNNs are good models of visual processing, as the performance of the DNNs can be achieved using more biologically plausible mechanisms.

The approach in understanding visual processing in the ventral stream, is by taking apart the process into smaller subprocesses. Thus, for the higher-level vision, such as face recognition, the data these networks can be trained on is still limited. Therefore, a clear caveat is the lack of data to train and validate the networks necessary for research questions aimed at biological plausibility. However, more neuroscientific data is being collected that are more tailored towards such research questions and target areas involved in face perception. For example, Cao et al. (2022) recently published a large human single-neuron dataset for this purpose.

Testing a ground truth

Although these approaches can provide an interesting metric, it must be noted that they are focused on brain activity, as the name implies. However, it is not always clear how brain activity and behavioural data relate to each other. Additionally, although exploring the intermediate activations of networks and comparing what happens inside the neural nets to brain processes make it more likely that there are some similar mechanisms used, it does not provide insight into the performance of these networks and how human-like that performance is. Lastly, and most important, it appears that the methods might not be as sensitive to assessing brain-likeness as we might assume (Han et al., 2022).

One of the most common approaches in computational neuroscience in evaluating brain-likeness, is to compare network outputs and brain activity. If there is a strong relationship or good predictability, this is assumed to mean validation of the model. In an elegantly thought up experiment, Han et al. (2022) set out to assess this method for neural regression and centered kernel alignment, two commonly used evaluation techniques. Instead of comparing models to the brain, which is considered a black box, a DNN was selected to serve as ‘ground truth’ and other models compared to it. The main question was to assess whether the two methods would identify the correct model if compared to multiple models. They found that the correct identification of the ‘ground truth’ model was highly variable. Two reasons for this were identified, properties of the architecture of the ‘ground truth’ model and the stimulus data used to evaluate the models. Importantly, the greater the extent to which the stimuli used in the evaluation of the model reflects the training data of the ‘ground truth’ model, the higher the likelihood of the correct identification of the ‘ground truth’ model. Thus, understanding visual stimuli and understanding what is used to learn faces, is an important goal in the pursuit of identifying brain-like networks.

The main take-away from the emerging research on frameworks such as the brain-score, the brain-hierarchy score and the exploration of biologically plausible mechanisms should be difficulty in assessing the ground truth. Although Han et al (2022) can show that the current methods are lacking, there is no clear solution offered as this is an open challenge. Although only providing a starting point in this thesis, I hope it is clear that the addition from behavioural metrics can be a solution to some extent. Although some exist, the current metrics do not include tasks that are aimed at specific sub-processes of human face processing (tasks such as 2-AFC object categorization). The methods and findings presented in this thesis can be used to develop more specific tasks and could help understand brain-likeness in a much more human-like way.

10.4. Brief summary of the recent changes in the field face identity recognition

In the following section, I will provide a brief summary of the recent changes in the field of face identity recognition. I would like to return to this quote from Chapter 1 (section 1.4): “Although differences between familiar and unfamiliar face processing are evident, defining qualitatively what is different is difficult and has been researched to a limited degree.” What is missing from this quote is an indication of how long these questions already have provided a challenge to

researchers. Although difficult to pinpoint a certain starting point, a good indicator of the perpetuation of these challenges can be found in Chapter 12 in the book ‘Computational, geometric, and process perspectives on facial cognition: Contexts and challenges.’ During my PhD, I have in particular thought often about the following fragment: “Consider the following list of some of the questions currently being asked by face perception researchers: 1. Are faces special or does their perception only represent a special case of more generalized form perception? ... 11. Are faces that are familiar perceived better than novel ones? ... 16. How are faces represented in memory?” (Uttal, 2001). For the attentive reader, these were the current questions as published in a book over 20 years ago. Yet, this thesis is concerned with the similar questions. Although a lot of progress has been made in the field, the perpetuation of these open questions should be taken as an indication of the immense challenge involved in understanding face identity recognition.

As discussed in Chapter 7, recent technological advances are likely to be related to the current renewed uprise of mathematical psychology that led to the growing field of computational cognition. As should be evident from the described future directions, the next step is to develop an integrative approach that focuses on exploring face perception through computational experiments in order to unify our understanding of previously collected data. The major challenge in this approach is to extract meaning from the wealth of information that are in DNNs. In the time of this PhD project, steps towards such an integrative approach have already been made, as presented in the discussed literature (Abudarham et al., 2021; Blauch et al., 2021a; Hancock et al., 2020; O’Toole & Castillo, 2021) and the research undertaken in this thesis. Each of the studies provide novel methods to use DNNs to gain new insights in (human) face perception.

I have stated that the most prominent challenge in my PhD was the technical skill required to implement new findings in computational neuroscience and techniques in computing science. However, it is clear that a computation approach is a promising direction to progress our understanding of human face recognition. The field of face identity recognition is clearly shifted to an interdisciplinary approach and it is important that future psychologists are trained in increasingly more interdisciplinary ways.

10.5. Conclusions

My research provides compelling groundwork and proposes a comprehensive and computational approach to explore the internal representation we hold of faces. This thesis can be summed up in four major findings: there is a linear relationship between image-based spaces and perceived similarity that can be used to explore the effect of PCA transformations, familiar faces are perceived as more similar overall compared to unfamiliar faces, the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces, and that the effect of familiarity can be found at the representational level (in DNNs). Although many questions remain open, it should be evident from this thesis that a computational cognitive approach to face recognition, although technically challenging and on many occasions laborious, will progress our understanding of human face recognition. And, with all the (technological) progress and increasing interdisciplinarity in face recognition, it should be clear that the time for this, is now.

Supplementary materials

Supplementary material – Chapter 3

Table 8 Distribution of familiarity in experiment 4a.

ID number	Percentage familiar	Percentage unfamiliar
1	67.2	32.8
2	21.3	78.7
3	62.3	37.7
4	68.9	31.1
5	16.4	83.6
6	31.1	68.9
7	73	39.3
8	72.1	27.9

Table 9 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 4a.

Predictors	Estimates	SE	df	t	p
(intercept)	5.11	.10	110.89	49.46	<.001***
Transformation Level 3	-1.05	.09	1584.37	-11.83	<.001***
Transformation Level 5	-1.86	.09	1665.97	-20.60	<.001***
Transformation Level 7	-2.00	.09	1527.82	-22.48	<.001***
Familiarity	.33	.13	630.66	2.59	.010**
Transformation Level 3 x Familiarity	.17	.19	200.70	.92	.359
Transformation Level 5 x Familiarity	.10	.18	287.58	.56	.574
Transformation Level 7 x Familiarity	.04	.19	167.31	.20	.843

Note * $p < .05$; ** $p < .01$; *** $p < .001$

Table 10 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 4a.

Predictors	Estimates	SE	df	t	p
(intercept)	4.06	.10	116.12	38.83	<.001***
Transformation Level 2	1.06	.09	1584.37	11.83	<.001***
Transformation Level 5	-.81	.09	1794.75	-9.11	<.001***
Transformation Level 7	-.95	.09	1833.78	-10.74	<.001***
Familiarity	.51	.15	79.49	3.35	.001**
Transformation Level 3 x Familiarity	-.17	.19	200.70	-.92	.359
Transformation Level 5 x Familiarity	-.07	.18	947.76	-.39	.695
Transformation Level 7 x Familiarity	-.14	.18	1798.11	-.77	.444

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

Table 11 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 4a.

Predictors	Estimates	SE	df	t	p
(intercept)	3.25	.10	112.56	31.26	<.001***
Transformation Level 2	1.86	.09	1665.97	20.96	<.001***
Transformation Level 3	.81	.09	1794.75	9.11	<.001***
Transformation Level 7	-.14	.09	1813.79	-1.61	.11
Familiarity	.44	.14	114.138	3.06	.002**
Transformation Level 2 x Familiarity	-.10	.18	287.58	-.56	.575
Transformation Level 3 x Familiarity	.07	.18	947.76	.39	.695
Transformation Level 7 x Familiarity	-.07	.18	1146.43	-.37	.712

Note: * p<.05; ** p<.01; *** p<.001

Supplementary materials – Chapter 4

Table 12 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 4b.

Predictors	Estimates	SE	df	t	p
(intercept)	5.03	0.12	60.77	43.45	<.001***
Transformation Level 3	-0.19	0.08	98.66	-2.39	.019**
Transformation Level 5	-0.44	0.08	234.18	-5.51	<.001***
Transformation Level 7	-0.77	0.09	65.31	-8.26	<.001***
Familiarity	0.20	0.11	843.02	1.77	.077
Transformation Level 3 x Familiarity	0.21	0.17	124.43	1.29	.201
Transformation Level 5 x Familiarity	-0.03	0.18	90.74	-0.16	.872
Transformation Level 7 x Familiarity	0.09	0.16	189.53	0.54	.590

Note * $p < .05$; ** $p < .01$; *** $p < .001$

Table 13 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 4b.

Predictors	Estimates	SE	df	t	p
(intercept)	4.83	0.11	61.93	43.60	<.001***
Transformation Level 2	0.19	0.08	98.90	2.39	.019*
Transformation Level 5	-0.25	0.08	250.96	-3.16	.002**
Transformation Level 7	-0.58	0.09	76.68	-6.40	<.001***
Familiarity	0.41	0.13	82.59	3.19	.002**
Transformation Level 3 x Familiarity	-0.21	0.17	124.40	-1.29	0.201
Transformation Level 5 x Familiarity	-0.24	0.16	153.99	-1.55	0.124
Transformation Level 7 x Familiarity	-0.13	0.16	122.11	-0.80	0.424

Note * $p < .05$; ** $p < .01$; *** $p < .001$

Table 14 Fixed effect of the linear mixed effects model for familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 4b.

Predictors	Estimates	SE	df	t	p
(intercept)	4.59	0.12	61.46	37.00	<.001***
Transformation Level 2	0.44	0.08	233.19	5.51	<.001***
Transformation Level 3	0.25	0.08	250.54	3.16	.002**
Transformation Level 7	-0.33	0.09	75.82	-3.68	<.001***
Familiarity	0.17	0.14	68.07	1.18	0.243
Transformation Level 2 x Familiarity	0.03	0.18	90.63	0.16	0.872
Transformation Level 3 x Familiarity	0.24	0.16	153.35	1.55	0.124
Transformation Level 7 x Familiarity	0.12	0.16	203.82	0.72	0.470

Note: * $p < .05$; ** $p < .01$; *** $p < .001$

Supplementary materials – Chapter 8

Table 15 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 7a.

Predictors	Estimates	SE	df	t	p
(intercept)	2.19	.06	30.01	34.62	<.001***
Transformation Level 3	.44	.02	38.64	19.21	<.001***
Transformation Level 5	1.01	.03	30.01	30.08	<.001***
Transformation Level 7	1.07	.04	30.25	26.95	<.001***
Familiarity	.01	.03	158.41	0.45	.652
Transformation Level 3 x Familiarity	-.02	.04	232.28	-.66	.510
Transformation Level 5 x Familiarity	-.02	.037	115.31	-.73	.457
Transformation Level 7 x Familiarity	-.02	.036	137.19	-.51	.614

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

Table 16 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 3 as reference for Experiment 7a.

Predictors	Estimates	SE	df	t	p
(intercept)	2.64	.08	30.01	34.80	<.001***
Transformation Level 2	-.44	.02	38.64	-19.21	<.001***
Transformation Level 5	.57	.03	34.06	22.09	<.001***
Transformation Level 7	.63	.04	30.11	17.06	<.001***
Familiarity	-.01	.03	53.56	-.41	.68
Transformation Level 2 x Familiarity	.02	.04	232.3	.66	.51
Transformation Level 5 x Familiarity	-.003	.03	1065	-.09	.925
Transformation Level 7 x Familiarity	.01	.03	2244	.15	.881

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

Table 17 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 7a.

Predictors	Estimates	SE	df	t	p
(intercept)	3.21	.08	30	39.32	<.001***
Transformation					<.001***
Level 2	-.57	.03	34.06	-22.09	
Transformation					<.001***
Level 3	-1.02	.03	30.01	-30.08	
Transformation					
Level 7	.05	.03	31.67	2.14	0.0402*
Familiarity	-.02	.03	43.36	-.49	.6254
Transformation					
Level 2 x					
Familiarity	.003	.03	1065	.09	.925
Transformation					
Level 3 x					
Familiarity	.03	.04	115.3	.73	.4667
Transformation					
Level 7 x					
Familiarity	.01	.03	3025	.25	.8065

Note: * p<.05; ** p<.01; *** p<.001

Table 18 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level in Euclidean distance with Transformation Level 2 as reference for Experiment 7b.

Predictors	Estimates	SE	df	t	p
(intercept)	2.92E-01	1.63E-03	90.52	179.88	<.001***
Transformation					<.001***
Level 3	3.62E-02	2.11E-03	4094	17.15	
Transformation					<.001***
Level 5	7.95E-02	2.11E-03	1770	37.60	
Transformation					<.001***
Level 7	1.13E-01	2.12E-03	1001	53.08	
Familiarity	-3.61E-03	3.37E-03	70.29	-1.07	0.288
Transformation					
Level 3 x					
Familiarity	8.39E-04	4.24E-03	1154	0.20	0.843
Transformation					
Level 5 x					
Familiarity	6.94E-03	4.23E-03	1695	1.64	.101
Transformation					
Level 7 x					
Familiarity	1.09E-02	4.29E-03	356.4	2.55	.011**

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

Table 19 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 2 as reference for Experiment 7c.

Predictors	Estimates	SE	df	t	p
(intercept)	1.53E-02	8.58E-04	30.02	17.81	<.001***
Transformation					<.001***
Level 3	9.72E-03	7.60E-04	30.05	12.79	
Transformation					<.001***
Level 5	1.36E-02	1.02E-03	29.98	13.39	
Transformation					<.001***
Level 7	1.87E-02	1.07E-03	29.99	17.51	
Familiarity	2.67E-04	3.51E-04	82.73	0.76	0.45
Transformation					
Level 3 x					
Familiarity	-6.09E-04	4.90E-04	92.58	-1.25	0.216
Transformation					
Level 5 x					
Familiarity	-2.70E-04	4.91E-04	89.04	-0.55	0.583
Transformation					
Level 7 x					
Familiarity	-3.29E-04	7.55E-04	32.92	-0.44	0.666

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

Table 20 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 5 as reference for Experiment 7c.

Predictors	Estimates	SE	df	t	p
(intercept)	2.89E-02	1.52E-03	3.00E+01	19.03	<.001***
Transformation					<.001***
Level 2	-1.36E-02	1.02E-03	3.00E+01	-13.39	
Transformation					<.001***
Level 3	-3.88E-03	5.28E-04	3.05E+01	-7.35	
Transformation					<.001***
Level 7	5.12E-03	5.34E-04	3.03E+01	9.58	
Familiarity	-3.75E-06	4.00E-04	4.37E+01	-0.01	0.993
Transformation					
Level 2 x					
Familiarity	2.70E-04	4.91E-04	8.91E+01	0.55	0.583
Transformation					
Level 3 x					
Familiarity	-3.39E-04	4.51E-04	8.85E+02	-0.75	0.453
Transformation					
Level 7 x					
Familiarity	-5.84E-05	6.23E-04	3.89E+01	-0.09	0.926

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

Table 21 Fixed effect of the linear mixed effects model for Familiarity and Transformation Level with Transformation Level 7 as reference for Experiment 7c.

Predictors	Estimates	SE	df	t	p
(intercept)	3.40E-02	1.57E-03	3.00E+01	21.69	<.001***
Transformation					<.001***
Level 2	-1.87E-02	1.07E-03	3.00E+01	-17.52	
Transformation					<.001***
Level 3	-9.00E-03	8.32E-04	3.00E+01	-10.82	
Transformation					
Level 5	-5.12E-03	5.34E-04	3.03E+01	-9.58	<.001***
Familiarity	-6.21E-05	6.47E-04	3.07E+01	-0.10	0.924
Transformation					
Level 2 x					
Familiarity	3.29E-04	7.55E-04	3.29E+01	0.44	0.666
Transformation					
Level 3 x					
Familiarity	-2.81E-04	6.21E-04	4.03E+01	-0.45	0.654
Transformation					
Level 5 x					
Familiarity	5.84E-05	6.23E-04	3.89E+01	0.09	0.926

Note: * p<.05; ** p<.01; *** p<.001

Table 22 Performance of the models in Experiment 7a

Model	Loss last epoch	Accuracy last epoch	Loss evaluation	Accuracy evaluation
1a	0.0113	1	0.8972	0.7708
1b	0.0123	0.9997	0.9748	0.7328
2a	0.0126	0.9997	0.9233	0.7647
2b	0.0141	0.9991	1.0078	0.7304
3a	0.0133	0.9997	0.9906	0.7574
3b	0.0126	0.9991	0.986	0.7439
4a	0.0132	1	0.986	0.7623
4b	0.0117	1	1.0242	0.7353
5a	0.0136	0.9994	0.9953	0.739
5b	0.012	1	1.0609	0.7316
6a	0.0131	0.9997	0.9683	0.7525
6b	0.0129	0.9994	1.0411	0.7402
7a	0.0129	1	1.138	0.7083
7b	0.013	1	0.8897	0.7525
8a	0.0112	0.9997	1.0022	0.7525
8b	0.0128	0.9997	1.0226	0.7586
9a	0.0135	0.9997	1.0597	0.7218
9b	0.0126	0.9997	1.015	0.7463
10a	0.0117	0.9997	0.9368	0.7586
10b	0.0115	1	0.9797	0.75
11a	0.0136	0.9994	0.9421	0.7574
11b	0.0136	0.9997	0.9609	0.7512
12a	0.0115	0.9997	0.9348	0.7708
12b	0.0123	0.9997	1.0091	0.7328
13a	0.0126	0.9997	0.8444	0.761
13b	0.013	0.9997	1.0362	0.7439
14a	0.0129	1	1.0026	0.7279
14b	0.0128	0.9997	0.9683	0.7549
15a	0.0125	1	0.9385	0.7537
15b	0.0124	0.9997	1.0769	0.7353

Table 23 Performance of the models in Experiment 7b

Model	Loss last epoch	Accuracy last epoch	Loss evaluation	Accuracy evaluation
1a	0.0021	0.9994	0.018	0.9951
1b	0.0025	0.9994	0.0139	0.9951
2a	0.0036	0.9997	0.0192	0.9939
2b	0.0037	0.9982	0.0181	0.9951
3a	0.0052	0.9985	0.0113	0.9939
3b	0.0044	0.9991	0.0369	0.9914
4a	0.0021	0.9997	0.0074	0.9951
4b	0.0025	0.9991	0.0394	0.9939
5a	0.0022	0.9997	0.0215	0.9951
5b	0.005	0.9976	0.0164	0.9926
6a	0.0047	0.9982	0.0137	0.9963
6b	0.0023	0.9994	0.0139	0.9951
7a	0.0037	0.9988	0.0191	0.9926
7b	0.0025	0.9991	0.0086	0.9963
8a	0.0035	0.9985	0.015	0.9963
8b	0.0032	0.9988	0.0133	0.9951
9a	0.0028	0.9991	0.0359	0.9926
9b	0.003	0.9991	0.0111	0.9939
10a	0.0039	0.9985	0.0143	0.9975
10b	0.0036	0.9991	0.019	0.9951
11a	0.0036	0.9988	0.013	0.9939
11b	0.0018	0.9997	0.0129	0.9951
12a	0.0021	0.9994	0.0291	0.9951
12b	0.0027	0.9988	0.0103	0.9963
13a	0.0019	1	0.0137	0.9939
13b	0.0021	0.9997	0.0274	0.9914
14a	0.0047	0.9976	0.006	0.9963
14b	0.0032	0.9994	0.0359	0.9914
15a	0.0046	0.9985	0.0154	0.9951
15b	0.0035	0.9982	0.0178	0.9939

Table 24 Performance of the models in Experiment 7c

Model	Loss last epoch	Accuracy last epoch	Loss evaluation	Accuracy evaluation
1a	0.0048	1	0.009	0.9988
1b	0.0055	1	0.0197	0.9975
2a	0.0053	1	0.0152	0.9988
2b	0.0055	1	0.0146	0.9963
3a	0.0056	1	0.0152	0.9988
3b	0.0056	1	0.0141	0.9975
4a	0.0053	1	0.021	0.9963
4b	0.0051	1	0.0097	1
5a	0.0056	1	0.0139	0.9988
5b	0.0053	1	0.0103	0.9975
6a	0.0056	1	0.0216	0.9963
6b	0.0055	1	0.0116	0.9988
7a	0.0055	1	0.0115	0.9963
7b	0.0056	1	0.0194	0.9963
8a	0.0055	1	0.0187	0.9963
8b	0.0052	1	0.0099	0.9963
9a	0.0051	1	0.0135	0.9963
9b	0.0061	0.9997	0.0146	0.9975
10a	0.0054	1	0.0086	0.9988
10b	0.0053	1	0.0156	0.9988
11a	0.0052	1	0.0088	0.9988
11b	0.0055	1	0.0234	0.9951
12a	0.0055	1	0.0107	0.9988
12b	0.0052	1	0.0308	0.9926
13a	0.0048	1	0.0115	0.9963
13b	0.0058	1	0.0201	0.9951
14a	0.0053	1	0.0195	0.9963
14b	0.0056	1	0.0199	0.9975
15a	0.005	1	0.0193	0.9963
15b	0.0053	1	0.0218	0.9939

Supplementary materials – Chapter 9

Table 25 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in cosine similarity, with Transformation Level 2 as reference for Experiment 7b.

Predictors	Estimates	SE	df	t	p
(intercept)	1.21E-01	1.61E-03	5731	75.43	<.001***
Transformation Level 3	3.68E-02	2.27E-03	5730	16.19	<.001***
Transformation Level 5	8.88E-02	2.27E-03	5730	39.09	<.001***
Transformation Level 7	1.21E-01	2.27E-03	5730	53.22	<.001***
Familiarity	-3.48E-03	3.63E-03	76.21	-0.96	0.34
Transformation Level 3 x Familiarity	-1.49E-03	4.57E-03	1066	-0.33	0.744
Transformation Level 5 x Familiarity	1.20E-03	4.59E-03	681.2	0.26	0.794
Transformation Level 7 x Familiarity	6.23E-03	4.60E-03	605.4	1.36	0.176

Note * $p < .05$; ** $p < .01$; *** $p < .001$

Table 26 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Manhattan distance, with Transformation Level 2 as reference for Experiment 7b.

Predictors	Estimates	SE	df	t	p
(intercept)	11.70	6.39E-02	92.89	183.07	<.001***
Transformation Level 3	1.44	8.29E-02	5420.13	17.34	<.001***
Transformation Level 5	3.13	8.29E-02	4332.28	37.76	<.001***
Transformation Level 7	4.42	8.31E-02	1780.14	53.16	<.001***
Familiarity	-0.16	1.31E-01	77.15	-1.23	0.222
Transformation Level 3 x Familiarity	0.03	1.67E-01	1131.31	0.19	0.853
Transformation Level 5 x Familiarity	0.30	1.66E-01	2000.86	1.81	0.071
Transformation Level 7 x Familiarity	0.46	1.68E-01	449.82	2.74	.006**

Note * $p < .05$; ** $p < .01$; *** $p < .001$

Table 27 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Jaccard-Global distance, with Transformation Level 2 as reference for Experiment 7b.

Predictors	Estimates	SE	df	t	p
(intercept)	5.19E-01	1.39E-03	388.6	373.79	<.001***
Transformation Level 3	4.22E-02	1.93E-03	3364	21.86	<.001***
Transformation Level 5	8.41E-02	1.93E-03	2030	43.48	<.001***
Transformation Level 7	1.08E-01	1.93E-03	5510	56.19	<.001***
Familiarity	-5.93E-03	3.30E-03	51.85	-1.80	0.078
Transformation Level 3 x Familiarity	1.74E-04	3.91E-03	567.5	0.05	0.965
Transformation Level 5 x Familiarity	8.20E-03	3.86E-03	5728	2.12	.034*
Transformation Level 7 x Familiarity	1.05E-02	3.87E-03	1381	2.70	0.006**

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

Table 28 Fixed effect of the linear mixed effects model for familiarity and Transformation Level measures in Jaccard-Detail distance, with Transformation Level 2 as reference for Experiment 7b.

Predictors	Estimates	SE	df	t	P
(intercept)	9.25E-01	3.43E-04	64.83	2694.89	<.001***
Transformation Level 3	7.16E-03	4.45E-04	507.7	16.08	<.001***
Transformation Level 5	1.47E-02	4.45E-04	555.9	32.96	<.001***
Transformation Level 7	1.82E-02	4.43E-04	794.6	41.10	<.001***
Familiarity	-1.17E-03	7.35E-04	55.8	-1.59	0.119
Transformation Level 3 x Familiarity	-1.11E-04	9.05E-04	237.9	-0.12	0.902
Transformation Level 5 x Familiarity	1.83E-03	8.81E-04	3070	2.08	.038*
Transformation Level 7 x Familiarity	2.63E-03	8.81E-04	2083	2.98	.003**

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

Supplementary materials – General discussion

Table 29 Overview of findings

Ch.	Title	Experiments	Findings
1	Chapter 1 - Introduction - the psychology of face representations		
2	Chapter 2 - Beyond binary face recognition task: The effects of familiarity on sensitivity to subtle face changes	Experiment 1: Effects of familiarity on sensitivity to two types of adjustments were studied, eyebrow and lip thickness (high perceptual sensitivity according to Abudarham et al. 2019) and eye distance and mouth width (low perceptual sensitivity). this first experiment had an on-screen comparison (face perception) Experiment 2: The second experiment was the same, bar the presentation of the stimuli, which was a delayed comparison (face recognition).	Experiment 1 and 2: Our findings indicate that perceptual sensitivity but not familiarity has an influence on the accuracy of our face perception, but neither influence the accuracy of recall of the selected manipulations.
3	Chapter 3 - Exploring perceptual similarity and its relation to image-based spaces: an effect of familiarity.	Experiment 3: we explore whether PCA could be used to create controlled transformations to a face by testing the effect of these (single component) transformations on human perceptual similarity and on computational differences in Gabor, Pixel and DNN spaces. Experiment 4a: The effect of familiarity on the perception of multidimensional transformations was explored.	Experiment 3: We found that perceptual similarity and the three image-based spaces are linearly related. Experiment 4a: Our findings show that there is a positive relationship between the number of components transformed and both the perceptual similarity and the same three image-based spaces used in experiment 1. Furthermore, we found that familiar faces are rated more similar overall than unfamiliar faces.

Ch.	Title	Experiments	Findings
4	Chapter 4 - Exploring other aspects of perceptibility of change: magnitude of change	Experiment 4b: I explored if the pattern found in Experiment 4a would be the same when the changes were smaller (2 SD instead of 4 SD) in terms of magnitude of to better understand the relationship between the number of components and magnitude of change, and to explore if the pattern we found in experiment 4a would be the same.	Experiment 4b: Our findings show a similar result as in experiment 4a, but, with higher overall similarity scores, and the difference between changes with 5 and 7 components is significant.
5	Chapter 5 - Title: Familiarity and quantifying sensitivity to dissimilarity	Experiment 5a: A 2-alternative forced choice task was used to explore differences between familiar and unfamiliar faces in the ability to identify changes made to the appearance of the face. The task was carried out with transformation of 2SD changes. Experiment 5b: The same experiment was run with cross-identity changes, 4 SD changes.	Experiment 5a and 5b: A main effect of perceptibility of change (in number of components) and Familiarity between the response pairs was found. Additionally, I found an interaction between familiarity and perceptibility of change. This means that the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces. Both experiments revealed a similar pattern of results.

Ch.	Title	Experiments	Findings
6	Chapter 6 - Finding the category boundaries of faces: familiarity and the distinctiveness of transformations	I explored the effect of varying the principal components used in the transformations on perceived identity (distinctiveness of transformation). A 4-AFC paradigm was explored that compares the effect of different principal components and test whether it can be used to select more distinctive PCA components. The secondary aim of this chapter is to further explore the effect of familiarity on sensitivity to change. The experiment was carried out twice, with different sets of identities: Experiment 6a used the PCA space used in the previous experiments, and Experiment 6b a new PCA space containing 4 new identities (4 identities replaced compared to the previous PCA space).	Experiment 6a: The results of Experiment 6a in relation to familiarity were difficult to interpret due to a large majority being familiar and unfamiliar with the same identities. There appeared to be a preference for a certain component dependent on the identity, one that cannot be reasoned away by the size of the change measured in an image-based space (Gabor distance). Experiment 6b: The results of Experiment 6b show that component 12 has the strongest effect on appearance in all identities, though the preference is minor compared to the effects of components seen in Experiment 6a. Since the same component affected all identities in a similar fashion, there were no clear indicators that familiarity influenced the distinctiveness of transformation.
7	Chapter 7 - Review on the use of deep neural networks in understanding human face recognition		

Ch.	Title	Experiments	Findings
8	Chapter 8 - Exploring the representation of familiar faces using deep neural networks.	A method of familiarization of faces in the form of fine-tuning of a pre-trained DNN is used to explore the two main findings presented in the thesis: familiar faces are perceived as more similar overall than unfamiliar faces (Experiment 4a), and the increase in sensitivity for greater changes between pairs is larger for familiar faces than for unfamiliar faces (Experiment 5a and 5b). Three experiments were run: Experiment 7a: the FaceNet architecture was used with 341 layers frozen, and 106 layers trained. Experiment 7b: the VGGFace architecture was used with 30 layers frozen and 8 layers trained. Experiment 7c: The model in Experiment 7a was retrained without any frozen layers to achieve maximum performance on the test set.	Experiment 7a: No effects of familiarity were found. Experiment 7b: A main effect for both Transformation Level and Familiarity was found. Additionally, there was a significant interaction between the two. Experiment 7c: Similar to Experiment 7a, no effect of Familiarity or significant interaction between Familiarity and Transformation Level were found in Experiment 7c, there was only a main effect of Transformation Level.

9	Chapter 9 - The relationship between distance measures of representational space and behavioural data	I further explored the results of Experiment 7b and compared the representational spaces using different distance measures, to find which outcome measure best reflects a human behavioural pattern.	Cosine, Manhattan and Jaccard-Global distance all showed the same pattern of results as found for the Euclidean distance in Experiment 7b: a significant main effect for Transformation Level but not for Familiarity, and a significant interaction effect between Transformation Level and Familiarity for Transformation Level 5 and 7. The pattern of results for the Jaccard-detail measure was almost as similar as the other measures, a significant main effect for Transformation Level but not for Familiarity, but only a significant interaction effect between Transformation Level and Familiarity for Transformation 7 and not for Transformation Level 5.
10	Chapter 10 - Conclusions and future directions		

Glossary

Term	Definition
ANN	Artificial Neural Network
DL	Deep Learning
DNN	Deep Neural Network
FIE	Face Inversion Effect
fMRI	Functional Magnetic Resonance Imaging
FRU	Face Recognition Unit
GPU	Graphics Processing Units
IAC	Interactive Activation and Competition (model)
ICA	Independent Components Analysis
ORE	Other-race effect
PCA	Principal Components Analysis
PS	Perceptual Sensitivity
RBF	Radial Basis Function
SWM	Spatial Working Memory
VWM	Visual Working Memory

References

- 't Hart, B. M., Abresch, T. G. J., & Einhäuser, W. (2011). Faces in Places : Humans and Machines Make Similar Face Detection Errors. *PLoS ONE*, 6(10), e25373. <https://doi.org/10.1371/journal.pone.0025373>
- Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *Cognitive Science*, 45(9), e13031. <https://doi.org/10.1111/cogs.13031>
- Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition*, 182, 73–83.
- Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, 16(3), 40. <https://doi.org/10.1167/16.3.40>
- Abudarham, N., & Yovel, G. (2018). Same critical features are used for identification of familiarized and unfamiliar faces. *Vision Research*, 157, 105–111. <https://doi.org/10.1016/j.visres.2018.01.002>
- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. *International Conference on Database Theory, January*, 420–434.
- Angeli, A., Davidoff, J., & Valentine, T. (2008). Face familiarity, distinctiveness, and categorical perception. *Quarterly Journal of Experimental Psychology*, 61(5), 690–707. <https://doi.org/10.1080/17470210701399305>
- Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, 57(3), 217–239. [https://doi.org/10.1016/0010-0277\(95\)00669-X](https://doi.org/10.1016/0010-0277(95)00669-X)
- Besson, G., Barragan-Jason, G., Thorpe, S. J., Fabre-Thorpe, M., Puma, S., Ceccaldi, M., & Barbeau, E. J. (2017). From face processing to face recognition: Comparing three different processing levels. *Cognition*, 158, 33–43. <https://doi.org/10.1016/j.cognition.2016.10.004>
- Bindemann, M., & Hole, G. J. (2020). Understanding face identification through within-person variability in appearance: Introduction to a virtual special issue. *Quarterly Journal of Experimental Psychology*, 73(12), NP1–NP8. <https://doi.org/10.1177/1747021820959068>
- Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021a). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, 208, 104341.

<https://doi.org/10.1016/j.cognition.2020.104341>

- Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021b). Deep learning of shared perceptual representations for familiar and unfamiliar faces: Reply to commentaries. *Cognition*, 208(August 2020), 104484. <https://doi.org/10.1016/j.cognition.2020.104484>
- Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2016). Eye-movement strategies in developmental prosopagnosia and “Super” face recognition. *The Quarterly Journal of Experimental Psychology*, 70(2), 201–217. <https://doi.org/10.1080/17470218.2016.1161059>
- Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. *Robustness in Statistics*, 201–236. <https://doi.org/10.1016/b978-0-12-438150-6.50018-2>
- Brooks, K. R., & Kemp, R. I. (2007). Sensitivity to feature displacement in familiar and unfamiliar faces: Beyond the internal/external feature distinction. *Perception*, 36(11), 1646–1659. <https://doi.org/10.1068/p5675>
- Broomhead, D. S., & Lowe, D. (1988). Royal Signals and Radar Establishment. In *Royal Signals and Radar Establishment Malvern (United Kingdom)*.
- Bruce, V. (1986). Influences of familiarity on the processing of faces. *Perception*, 15(4), 387–397. <https://doi.org/10.1068/p150387>
- Bruce, V., Dench, N., & Burton, M. (1993). Effects of distinctiveness, repetition and semantic priming on the recognition of face familiarity. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, 47(1), 38–60. <https://doi.org/10.1037/h0078773>
- Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(4), 339–360. <https://doi.org/10.1037/1076-898X.5.4.339>
- Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, 7(3), 207–218. <https://doi.org/10.1037/1076-898X.7.3.207>
- Bruce, V., & Langton, S. (1994). The use of pigmentation and shading information in recognising the sex and identities of faces. *Perception*, 23(7), 803–822. <https://doi.org/10.1068/p230803>
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305–327. <https://doi.org/10.1111/j.2044-8295.1986.tb02199.x>

- Burgess, N., & Hitch, G. (2005). Computational models of working memory: Putting long-term memory into context. *Trends in Cognitive Sciences*, 9(11), 535–541.
<https://doi.org/10.1016/j.tics.2005.09.011>
- Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, 23(1), 1–31. https://doi.org/10.1207/s15516709cog2301_1
- Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, 81, 361–380.
- Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, 51(3), 256–284.
<https://doi.org/10.1016/j.cogpsych.2005.06.003>
- Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, 102(4), 943–958. <https://doi.org/10.1111/j.2044-8295.2011.02039.x>
- Burton, A. M., Kramer, R. S. S., & Ritchie, K. L. (2016). Identity From Variation : Representations of Faces Derived From Multiple Instances. *Cognitive Science*, 40, 202–223.
<https://doi.org/10.1111/cogs.12231>
- Burton, A. M., & Vokey, J. R. (1998). The Face-Space Typicality Paradox: Understanding the Face - Space Metaphor. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 51(3), 475–483. <https://doi.org/10.1080/713755768>
- Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin and Review*, 11(4), 735–741.
<https://doi.org/10.3758/BF03196628>
- 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.
<https://doi.org/10.1162/089892905774597254>
- Cao, R., Lin, C., Brandmeir, N. J., & Wang, S. (2022). A human single-neuron dataset for face perception. *Scientific Data*, 9(1), 1–10. <https://doi.org/10.1038/s41597-022-01482-4>
- Carbon, C. C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, 37(5), 801–806. <https://doi.org/10.1068/p5789>

- Chaby, L., George, N., Renault, B., & Fiori, N. (2003). Age-related changes in brain responses to personally known faces: An event-related potential (ERP) study in humans. *Neuroscience Letters*, 349(2), 125–129. [https://doi.org/10.1016/S0304-3940\(03\)00800-0](https://doi.org/10.1016/S0304-3940(03)00800-0)
- Chance, J. E., Turner, A. L., & Goldstein, A. G. (1982). Development of differential recognition for own- and other-race faces. *The Journal of Psychology*, 112, 29–37.
- Christian, B. (2020). *The alignment problem*. Atlantic Books.
- Claypool, H. M., Hugenberg, K., Housley, M. K., & Mackie, D. M. (2007). Familiar eyes are smiling: On the role of familiarity in the perception of facial affect. *European Journal of Social Psychology*, 37, 856–866.
- Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, 31(8), 985–994. <https://doi.org/10.1068/p3335>
- Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, 11(7), 857–869. <https://doi.org/10.1080/13506280444000021>
- Corrow, S. L., Dalrymple, K. A., & Barton, J. J. S. (2016). Prosopagnosia : current perspectives. *Eye and Brain*, 8, 165–175.
- Craw, I. (1995). A manifold of face and object recognition. In *Cognitive and computational aspects of face recognition* (pp. 183–203).
- Craw, I., & Cameron, P. (1991). Parameterising images for recognition and reconstruction. *BMVC91*, 367–379. <https://doi.org/10.5244/c.5.52>
- Crookes, K., & Hayward, W. G. (2012). Face inversion disproportionately disrupts sensitivity to vertical over horizontal changes in eye position. *Journal of Experimental Psychology: Human Perception and Performance*, 38(6), 1428–1437. <https://doi.org/10.1037/a0027943>
- Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12, 1053–1073.
- De Valois, R. L., & De Valois, K. K. (1990). *Spatial vision*.
- 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. <https://doi.org/10.1016/j.neuropsychologia.2005.07.001>

- Dunn, J. D., Varela, V. P. L., Nicholls, V. I., Papinutto, M., White, D., & Miellet, S. (2021). Visual information sampling in super-recognizers. *PsyArXiv*, 1–13. <https://doi.org/10.31234/osf.io/z2k4a>
- Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing Higher-Layer Features of a Deep Network Technical Report 1341 The models*.
- Freiwald, W. A., Duchaine, B., & Yovel, G. (2017). Face Processing Systems. *Annual Review of Neuroscience*, 39, 325–346. <https://doi.org/10.1146/annurev-neuro-070815-013934>.FACE
- Friston, K. (2009). The free-energy principle : a rough guide to the brain ? *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Gauthier, I., & Tarr, M. J. (1997). Becoming a “Greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12), 1673–1682. [https://doi.org/10.1016/S0042-6989\(96\)00286-6](https://doi.org/10.1016/S0042-6989(96)00286-6)
- Goffaux, V., & Dakin, S. C. (2010). Horizontal information drives the behavioral signatures of face processing. *Frontiers in Psychology*, 1, 143. <https://doi.org/10.3389/fpsyg.2010.00143>
- Goffaux, V., & Rossion, B. (2007). Face Inversion Disproportionately Impairs the Perception of Vertical but not Horizontal Relations Between Features. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4), 995–1002. <https://doi.org/10.1037/0096-1523.33.4.995>
- Gosselin, F., & Schyns, P. G. (2001). Bubbles : a technique to reveal the use of information in recognition tasks. *Vision Research*, 41, 2261–2271.
- Han, Y., Poggio, T., Cheung, B., Han, Y., & Cheung, B. (2022). System identification of neural systems : If we got it right , would we know ? System identification of neural systems : If we got it right , would we know ? *PREPRINT*.
- Hancock, P. J. B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments, and Computers*, 32(2), 327–333. <https://doi.org/10.3758/BF03207802>
- Hancock, P. J. B. (2021). Familiar faces as islands of expertise. *Cognition*, 214(March), 104765. <https://doi.org/10.1016/j.cognition.2021.104765>
- Hancock, P. J. B., Bruce, V., & Burton, A. M. (1998). Testing Principal Component Representations for Faces. *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, 84–97. https://doi.org/10.1007/978-1-4471-1546-5_8

- Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, 4(9), 330–337. [https://doi.org/10.1016/S1364-6613\(00\)01519-9](https://doi.org/10.1016/S1364-6613(00)01519-9)
- Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition*, 24(1), 26–40. <https://doi.org/10.3758/BF03197270>
- Hancock, P. J. B., Burton, A. M., & Bruce, V. (1995). Preprocessing images of faces: correlations with human perceptions of distinctiveness and familiarity. *Fifth International Conference on Image Processing and Its Applications*, 727 – 731. <https://doi.org/10.1049/cp:19950755>
- Hancock, P. J. B., Somai, R. S., & Mileva, V. R. (2020). Convolutional neural net face recognition works in non-human-like ways. *Royal Society Open Science*, 7, 1–7.
- Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research*, 38(1), 125–137. [https://doi.org/10.1016/S0042-6989\(97\)00116-8](https://doi.org/10.1016/S0042-6989(97)00116-8)
- Helman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. A. (2018). Assessing the point at which averages are stable: A tool illustrated in the context of person perception. *PREPRINT*.
- Herzmann, G., Dantiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods*, 40(3), 840–857. <https://doi.org/10.3758/BRM.40.3.840>
- Hill, H., & Bruce, V. (1996). Effects of Lighting on the Perception of Facial Surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4), 986–1004. <https://doi.org/10.1037/0096-1523.22.4.986>
- Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition*, 62(2), 201–222. [https://doi.org/10.1016/S0010-0277\(96\)00785-8](https://doi.org/10.1016/S0010-0277(96)00785-8)
- Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(May), 1–15. <https://doi.org/10.1038/ncomms15037>
- Hox, J. (1998). *Multilevel Modeling: When and Why*. 1965, 147–154. https://doi.org/10.1007/978-3-642-72087-1_17
- Huang, G. B., Mattar, M., Berg, T., & Learned-miller, E. (2008). Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments. *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 7–49.
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face.

- Cognition*, 121(3), 313–323. <https://doi.org/10.1016/j.cognition.2011.08.001>
- Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. <https://doi.org/10.1109/cvprw.2009.5206848>
- Johnston, A., Hill, H., & Carman, N. (1992). Recognising faces- effects of lighting direction, inversion, and brightness reversal.pdf. *Perception*, 21, 365–375.
- Johnston, R. J., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, 17(5), 577–596. <https://doi.org/10.1080/09658210902976969>
- Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6), 1187–1211. <https://doi.org/10.1152/jn.1987.58.6.1187>
- Kämäräinen, J., Kyrki, V., & Kälviäinen, H. (2002a). Fundamental Frequency Gabor Filters. *Proceedings of ICPR2002 16th International Conference on Pattern Recognition*, 1, 628–631.
- Kämäräinen, J., Kyrki, V., & Kälviäinen, H. (2002b). *Gabor features for invariant object recognition*. Lappeenranta University of Technology.
- Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception*, 33(4), 399–408. <https://doi.org/10.1068/p5083>
- Kemp, R., Pike, G., White, P., & Musselman, A. (1996). Perception and recognition of normal and negative faces: The role of shape from shading and pigmentation cues. *Perception*, 25(1), 37–52. <https://doi.org/10.1068/p250037>
- Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1), 103–108.
- Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, 172(December 2017), 46–58. <https://doi.org/10.1016/j.cognition.2017.12.005>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>

- Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., Malsburg, C. von der, Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3), 300–311.
- Larson, C. L., Aronoff, J., & Stearns, J. J. (2007). The Shape of Threat : Simple Geometric Forms Evoke Rapid and Sustained Capture of Attention. *Emotion*, 7(3), 526–534. <https://doi.org/10.1037/1528-3542.7.3.526>
- Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 53(2), 513–536. <https://doi.org/10.1080/713755889>
- Leder, H., & Carbon, C. C. (2006). Face-specific configural processing of relational information. *British Journal of Psychology*, 97(1), 19–29. <https://doi.org/10.1348/000712605X54794>
- Lewis, M. B. (2004). Face-space-R: Towards a unified account of face recognition. In *Visual Cognition* (Vol. 11, Issue 1). <https://doi.org/10.1080/13506280344000194>
- Li, J., He, D., Zhou, L., Zhao, X., Zhao, T., Zhang, W., & He, X. (2019). The Effects of Facial Attractiveness and Familiarity on Facial Expression Recognition. *Frontiers in Psychology*, 10, 2496. <https://doi.org/10.3389/fpsyg.2019.02496>
- Lobue, V., Larson, C. L., Lobue, V., & Larson, C. L. (2010). *What makes an angry face look so ... angry ? Examining visual attention to the shape of threat in children and adults*. 6285. <https://doi.org/10.1080/13506281003783675>
- Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, 8(10), 1386–1390. <https://doi.org/10.1038/nn1538>
- Luck, S. J., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281. <https://doi.org/10.1038/36846>
- Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3), 347–356. <https://doi.org/10.1038/nn.3655>
- Malcolm, G. L., Leung, C., & Barton, J. J. S. (2004). Regional variation in the inversion effect for faces: Differential effects for feature shape, feature configuration, and external contour. *Perception*, 33(10), 1221–1231. <https://doi.org/10.1068/p5372>
- Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality*

- and *Social Psychology*, 13(4), 330–334. <https://doi.org/10.1037/h0028434>
- MATLAB. (2019). *MATLAB r2019a*. The Mathworks Inc.
- McClelland, J. L., & Rumelhart, D. E. (1980). *An Interactive Activation Model of the effect of context in perception*.
- McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. <https://doi.org/10.1038/264746a0>
- Mckone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? toward a new theory of holistic processing. *Psychonomic Bulletin and Review*, 16(5), 778–797. <https://doi.org/10.3758/PBR.16.5.778>
- Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, 34(4), 865–876.
- 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), 3–35. <https://doi.org/10.1037/1076-8971.7.1.3>
- Melnik, A., Schüler, F., Rothkopf, C. A., König, P., & Schmidt, J. C. (2018). The World as an External Memory : The Price of Saccades in a Sensorimotor Task. *Frontiers in Behavioral Neuroscience*, 12(November). <https://doi.org/10.3389/fnbeh.2018.00253>
- Menon, N., White, D., & Kemp, R. I. (2015). Identity-level representations affect unfamiliar face matching performance in sequential but not simultaneous tasks. *Quarterly Journal of Experimental Psychology*, 68(9), 1777–1793. <https://doi.org/10.1080/17470218.2014.990468>
- Meytlis, M. (2011). A model of face space. *Visual Cognition*, 19(1), 13–26. <https://doi.org/10.1080/13506285.2010.512507>
- Mike Burton, A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, 66(8), 1467–1485. <https://doi.org/10.1080/17470218.2013.800125>
- Mike Burton, A., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, 42(1), 286–291. <https://doi.org/10.3758/BRM.42.1.286>
- Mordvintsev, A., Tyka, M., & Olah, C. (2015). *DeepDream (google/deepdream) GitHub repository* Available at: <https://github.com/google/deepdream>.

- Moses, Y., Ullman, S., & Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception*, 25(4), 443–461. <https://doi.org/10.1068/p250443>
- Mousavi, S. M., & Oruc, I. (2020). Tuning of face expertise with a racially heterogeneous face-diet. *Visual Cognition*, 28(9), 523–539. <https://doi.org/10.1080/13506285.2020.1836696>
- Nishimura, M., Doyle, J., Humphreys, K., & Behrmann, M. (2010). Probing the face-space of individuals with prosopagnosia. *Neuropsychologia*, 48(6), 1828–1841. <https://doi.org/10.1016/j.neuropsychologia.2010.03.007>
- Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 103013. <https://doi.org/10.1016/j.isci.2021.103013>
- Noyes, E., & O'Toole, A. J. (2017). Face recognition assessments used in the study of super-recognizers. *ArXiv Preprint*, 1–6. <https://arxiv.org/pdf/1705.04739.pdf>
- O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, 30(6), 755–764. <https://doi.org/10.1068/p3027>
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience : Understanding the Mind by Simulating the Brain*.
- O'Toole, A. J., & Castillo, C. D. (2021). Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. *Annual Review of Vision Science*, 7, 543–570. <https://doi.org/10.1146/annurev-vision-093019-111701>
- O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition*, 26(1), 146–160. <https://doi.org/10.3758/BF03211378>
- O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, 38(15–16), 2351–2363. [https://doi.org/10.1016/S0042-6989\(98\)00042-X](https://doi.org/10.1016/S0042-6989(98)00042-X)
- Osborne, C. D., & Stevenage, S. V. (2008). Internal feature saliency as a marker of familiarity and configural processing. *Visual Cognition*, 16(1), 23–43. <https://doi.org/10.1080/13506280701238073>
- Palomares, J. K. S., & Young, A. W. (2018). Facial First Impressions of Partner Preference Traits : Trustworthiness , Status , and Attractiveness. *Social Psychological and Personality Science*, 9(8),

990–1000. <https://doi.org/10.1177/1948550617732388>

- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *British Machine Vision Conference*.
- Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory*, 3(4), 406–417. <https://doi.org/10.1037/0278-7393.3.4.406>
- Pogodin, R., Mehta, Y., Lillicrap, T. P., & Latham, P. E. (2021). Towards Biologically Plausible Convolutional Networks. *Advances in Neural Information Processing Systems*, 34, 13924–13936. <http://arxiv.org/abs/2106.13031>
- R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. <https://www.r-project.org/>
- Ramon, M. (2021). Super-Recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, 158, 107809. <https://doi.org/10.1016/j.neuropsychologia.2021.107809>
- Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 110(3), 461–479. <https://doi.org/10.1111/bjop.12368>
- Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition*, 26(3), 179–195. <https://doi.org/10.1080/13506285.2017.1405134>
- Rhodes, G. (1985). Lateralized processes in face recognition. *British Journal of Psychology*, 76, 249–271.
- Riesenhuber, M., Jarudi, I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proceedings of the Royal Society B: Biological Sciences*, 271(SUPPL. 6), 448–450. <https://doi.org/10.1098/rsbl.2004.0216>
- Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1), 455–463. <https://doi.org/10.1152/jn.2002.88.1.455>
- Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, 70(5), 897–905. <https://doi.org/10.1080/17470218.2015.1136656>
- Ritchie, K. L., Kramer, R. S. S., & Burton, A. M. (2018). What makes a face photo a ‘good likeness’? *Cognition*, 170, 1–8. <https://doi.org/10.1016/j.cognition.2017.09.001>

- Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2010). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, 144, 161–169. <https://doi.org/10.1016/j.actpsy.2011.12.005>
- Roberts, T., & Bruce, V. (1989). Repetition priming of face recognition in a serial choice reaction-time task. *British Journal of Psychology*, 80, 201–211.
<https://www.unhcr.org/publications/manuals/4d9352319/unhcr-protection-training-manual-european-border-entry-officials-2-legal.html?query=excom> 1989
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
- Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica*, 128(2), 274–289. <https://doi.org/10.1016/j.actpsy.2008.02.003>
- Rossion, B. (2018). Humans Are Visual Experts at Unfamiliar Face Recognition. *Trends in Cognitive Sciences*, 22(6), 471–472. <https://doi.org/10.1016/j.tics.2018.03.002>
- Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, 16(2), 252–257.
<https://doi.org/10.3758/PBR.16.2.252>
- Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception*, 32(3), 285–293. <https://doi.org/10.1068/p5027>
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. (2017). *Evaluating the Visualization of What a Deep Neural Network Has Learned*. 28(11), 2660–2673.
- Sandford, A., & Burton, A. M. (2014). Tolerance for distorted faces : Challenges to a configural processing account of familiar face recognition. *Cognition*, 132(3), 262–268.
<https://doi.org/10.1016/j.cognition.2014.04.005>
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, PREPRINT, 1–9. <https://doi.org/10.1101/407007>
- Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>

- Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, 13(5), 402–410.
<http://content.ebscohost.com/ContentServer.asp?EbscoContent=dGJyMNLe80Seqa44y9f3OLCmr1GeprVSrq%2B4SLKWxWXS&ContentCustomer=dGJyMPGptEy1qbJPuePfgeyx43zx1%2B6B&T=P&P=AN&S=R&D=a9h&K=7225333>
- Sengupta, B., Stemmler, M. B., & Friston, K. J. (2013). Information and Efficiency in the Nervous System — A Synthesis. *PLOS Computational Biology*, 9(7), 1–12.
<https://doi.org/10.1371/journal.pcbi.1003157>
- Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13–22.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.
- Somai, R. S., & Hancock, P. J. B. (2022). Exploring perceived face similarity and its relation to image-based spaces: an effect of familiarity. *Visual Cognition*, 21(9), 2149.
<https://doi.org/10.1167/jov.21.9.2149>
- Somai, R. S., Schut, M. J., & Van der Stigchel, S. (2019). Evidence for the world as an external memory : A trade-off between internal and external visual memory storage. *CORTEX*, 122, 108–114.
<https://doi.org/10.1016/j.cortex.2018.12.017>
- Susilo, T., McKone, E., Dennett, H., Darke, H., Palermo, R., Hall, A., Pidcock, M., Dawel, A., Jeffery, L., Wilson, C. E., & Rhodes, G. (2010). Face recognition impairments despite normal holistic processing and face space coding: Evidence from a case of developmental prosopagnosia. *Cognitive Neuropsychology*, 27(8), 636–664. <https://doi.org/10.1080/02643294.2011.613372>
- Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1701–1708. <https://doi.org/10.1109/CVPR.2014.220>
- Tappert, C. C. (2019). Who is the father of deep learning? *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, 343–348.
<https://doi.org/10.1109/CSCI49370.2019.00067>
- Tiddeman, B. (2011). Facial feature detection with 3D convex local models. *Face and Gesture 2011*,

- IEEE*(2011), 400–405. <https://doi.org/10.1109/FG.2011.5771433>
- Tiddeman, B., Burt, M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5), 42–50.
<https://doi.org/10.1109/38.946630>
- Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 71–86.
- Uttal, W. R. (2001). Are reductive (explanatory) theories of face identification possible? Some speculations and some findings. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*. (pp. 467–501). Lawrence Erlbaum Associates.
- Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. In *British Journal of Psychology* (Vol. 79, pp. 471–491).
- 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.
<https://doi.org/10.1080/14640749108400966>
- Valentine, T. (2001). Face-space models of face recognition. *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, July, 83–113.
<https://doi.org/10.1006/jmps.1997.1193>
- Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1. <https://doi.org/10.1109/cvpr.2001.990517>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272.
<https://doi.org/10.1038/s41592-019-0686-2>
- Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, 429, 215–244.
<https://doi.org/10.1016/j.neucom.2020.10.081>
- Wickham, L. H. V., Morris, P. E., & Fritz, C. O. (2000). Facial distinctiveness : Its measurement, distribution and influence on immediate and delayed recognition. *British Journal of Psychology*,

91(1), 99–123. <https://doi.org/10.1348/000712600161709>

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17(7), 592–598. <https://doi.org/10.1111/j.1467-9280.2006.01750.x>

Wiskott, L., Fellous, J. M., Krüger, N., & Von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 775–779. https://doi.org/10.1007/3-540-63460-6_150

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>

Young, A. W., & Burton, A. M. (2018a). Are We Face Experts? *Trends in Cognitive Sciences*, 22(2), 100–110. <https://doi.org/10.1016/j.tics.2017.11.007>

Young, A. W., & Burton, A. M. (2018b). What We See in Unfamiliar Faces: A Response to Rossion. *Trends in Cognitive Sciences*, 22(6), 472–473. <https://doi.org/10.1016/j.tics.2018.03.008>

Yue, X., Biederman, I., Mangini, M. C., Malsburg, C. von der, & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research*, 55, 41–46. <https://doi.org/10.1016/j.visres.2011.12.012>

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: Past , present and future. *Computer Vision and Image Understanding*, 138, 1–24. <https://doi.org/10.1016/j.cviu.2015.03.015>

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: a literature survey. *ACM Computing Surveys (CSUR)*, 35(4), 399–458. <https://doi.org/10.1109/TIP.2009.2021737>