Exploring the representation of caricatures, facial motion, and view-invariance in face space.

by

Ryan Elson

Thesis submitted for Degree of Doctor of Philosophy in Digital Society

December 2023

Abstract

Faces present a vast array of information, from invariable features such as identity, to variable features such as expression, speech and pose. Humans have an incredible capability of recognising faces (familiar faces at least) and interpreting facial actions, even across changes in view. While there has been an explosion of research into developing artificial neural networks for many aspects of face processing, some of which seem to predict neural responses quite well, the current work focuses on face processing through simpler linear projection spaces. These linear projection spaces are formal instantiations of ‘face space’, built using principal component analysis (PCA). The concept of ‘face space’ (Valentine, 1991) has been a highly influential account of how faces might be represented in the brain. In particular, recent research supports the presence of a face space in the macaque brain in the form of a linear projection space, referred to as ‘axis coding’ in which individual faces can be coded as linear sum of orthogonal features. Here, these linear projection spaces are used for two streams of investigation.

Firstly, we assessed the neurovascular response to hyper-caricatured faces in an fMRI study. Based on the assumption that faces further from average should project more strongly onto components in the linear space, we hypothesised that they should elicit a stronger response. Contrary to our expectations, we found little evidence for this in the fusiform face area (FFA) and face-selective cortex more generally, although the response pattern did become more consistent for caricatured faces in the FFA. We then explored the response to these caricatured faces in cortex typically associated with object processing. Interestingly, both the average response magnitude and response pattern consistency increased to these stimuli as caricaturing increased. At the current time it is unclear if this response allows some functional benefit for processing caricatured faces, or whether it simply reflects similarities in the low- and mid-level properties to certain objects. If the response is functional, then hyper-caricaturing could pave a route to improving face processing in individuals with prosopagnosia if technologies can be developed to automatically caricature faces in real-time.

The second line of work addressed these linear projection spaces in the context of achieving view-invariance, specifically in the domain of facial motion and expression. How humans create view-invariant representations is still of interest, despite much research, however little work has focused on creating view-invariant representations outside of identity recognition. Likewise, there has been much research into face space and view-invariance separately, yet there is little evidence for how different views may be represented within a face space framework, and how motion might also be incorporated.

Automatic face analysis systems mostly deal with pose by either aligning to a canonical frontal view or by using separate view-specific models. There is inconclusive evidence that the brain possesses an internal 3D model for ‘frontalising’ faces, therefore here we investigate how changes in view might be processed in a unified multi-view face space based on using a few prototypical 2D views. We investigate the functionality and biological plausibility of five identity-specific faces spaces, created using PCA, that allow for different views to be reconstructed from single-view video inputs of actors speaking. The most promising of these models first builds a separate orthogonal space for each viewpoint. The relationships between the components in neighbouring views are learned, and then reconstructions across views are made using a cascade of projection, transformation, and reconstruction. These reconstructions are then collated and used to build a multi-view space, which can reconstruct motion well across all learned views.

This provides initial insight into how a biologically plausible, view-invariant system for facial motion processing might be represented in the brain. Moreover, it also has the capacity to improve view-transformations in automatic lip-reading software.

Acknowledgements

First and foremost, I am extremely grateful to my supervisors, Prof Alan Johnston and Dr Denis Schluppeck for all their support and guidance. They encouraged me to develop my skills and try new things, and gave me the space to work as an independent researcher, all of which I am incredibly thankful for. I have learned a lot under their supervision from their wealth of experience and knowledge, and hope that soon I can pass their words of wisdom onto students of my own.

I also want to thank my fellow PhD students and colleagues for their friendship, support and academic advice. In particular, I would like to express my gratitude to Dr Nick Simonsen who helped with stimulus curation, advised me on methods and statistical analysis, and was always there when I needed to vent about MATLAB (or rather my code) not working, and Dr Karl Miller who was a font of knowledge about School and ESRC processes. I would also like to thank Dr Chris Scholes, Dr David Watson, Dr Ljubica Jovanovic, Dr Carlos Cassanello, Dr Dan Hu and Dr Kristian Skoczek for their assistance throughout. Also, Dr Fintan Nagle, for constructing the PCA spaces used in the fMRI study and for allowing us to share the spaces through the Open Science Framework. I’d also like to thank my participants, some of whom heroically sat through my 8-hour experiment.

I would also like to extend my appreciation to Dr David Pitcher at the University of York, who supervised my Master’s project, for his continued mentoring during my PhD and for bringing this opportunity to my attention. The team at the York Neuroimaging Centre also always welcomed me back whenever I fancied a change of scenery.

I am grateful to the Economic and Social Research Council for funding my PhD, for providing and encouraging additional training outside of my immediate field, and for the extension following the impact the Covid 19 pandemic. I am also thankful for the Researcher Academy who provided me with incredibly useful training and experiences during my PhD and the School of Psychology.

Finally, I would like to thank my wife, Jessica, and my family, particularly my parents and my mother-in-law, all of whom supported me through my PhD, through both the highs and the lows. Even when the PhD was overwhelming, they kept me going.

Contents

List of Figures

List of Tables

Publications

Elson, R., Schluppeck, D., and Johnston, A. (2023) fMRI evidence that hyper-caricatured faces activate object-selective cortex. *Frontiers in Psychology 13*. https://doi.org/10.3389/fpsyg.2022.1035524

I also contributed to the following paper during my PhD. This was not part of the PhD but is in the same line of research.

Johnston, A., Brown, B. B., & Elson, R. (2021) Synchronous facial action binds dynamic facial features. *Scientific Reports 11* (7191) https://doi.org/10.1038/s41598-021-86725-x

Conference abstracts

Elson, R., Valstar, M., Schluppeck, D., & Johnston, A. (2021, July). Viewing face space from a different angle. *I-PERCEPTION* 12(4), pp. 7-8. https://doi.org/10.1177/20416695211039080

Elson, R., Schluppeck, D., Valstar, M., & Johnston, A. (2022, April). Taking face space to the extreme: assessing the effect of hyper-caricaturing faces on the fMRI response in the FFA. *I-PERCEPTION* 51(5), pp. 354-364. https://doi.org/10.1177/03010066221091992

Elson, R., Schluppeck, D., and Johnston, A. (2023) Reconstructing facial motion across views using a multi-view face space. *Journal of Vision 23*(5453) https://doi.org/10.1167/jov.23.9.5453

Data availability

The generated datasets and experimental code for the fMRI study in Chapter 3 are available via the Open Science Framework (https://doi.org/10.17605/OSF.IO/JEC6U). The contents include the scripts for running the experiment, data analysis scripts, as well as relevant behavioural results and logs of the MRI scans. Raw MRI data is also included in the submission. The code and data for the remaining chapters is available upon request. Email ryan.elson.2019@gmail.com

Funding

This doctoral work was supported by the Economic and Social Research Council [grant number: ES/P000711/1]. The scan time for the fMRI study in Chapter 3 was awarded by the Sir Peter Mansfield Imaging Centre through a pump-priming scheme. Open access fees were supported by the UKRI block grant for the University of Nottingham.

Chapter 1 Introduction

1.1 Outline

The face is a highly informative piece of biological hardware, able to convey identity, expression, speech, age, gender and more. For familiar faces at least, humans have a remarkable capability for processing these factors under a huge range of visual variation. The complexity of this task is perhaps evidenced by face learning in humans taking >30 years to optimise (Germine et al., 2011).

It is estimated that on average, individuals have some degree of familiarity with circa 5000 faces (Jenkins et al., 2018), which means learning both the between-identity and within-identity variation for all of these faces.

Two of the major sources of within-identity variation come from changes in viewpoint and changes in dynamic expression such as during speech. Despite much research into face processing, there is still much to learn in the context of view-invariance and facial motion, particularly with respect to the concept of ‘face space’ (Valentine, 1991; Valentine et al., 2016).

Face space has been highly influential in both investigating neural representations in humans and non-human primates as well as in automatic face recognition systems. While it has seen utility in various aspects of face processing, relatively little research has addressed how face space can represent either different viewpoints or facial motion. Even less work has addressed view-invariant representations *of* facial motion.

Here we aim to address this gap in understanding. In this introduction, we will first outline what face space is and provide some examples of its uses. Then we will discuss what happens when face space is taken to the extreme through caricaturing. Subsequently, the discussion will turn to the problem of achieving view-invariance, drawing on the importance of facial motion, as well as evidence and mechanisms in macaques, humans and computational models.

1.2 Face space representations

1.2.1 Overview of face space representations

The theoretical account of face space (Valentine, 1991; Valentine et al., 2016) posits that faces sit within a multidimensional space. The origin represents the average of the faces we have encountered, and individual faces are represented either as exemplars at particular locations in the space, or as directions in the space. This latter explanation has been termed norm- or prototype-based encoding, with faces being processed relative to the origin.

The dimensions of this theoretical space can be combinations of shape and texture that are more abstract than the qualitative labels one might apply to discrete changes in single features. The closer to the origin a given face is on each dimension, the more average the features are. An identity's distinctiveness can be described by the direction in the space or the distance from average. If the dimensions are ordered hierarchically based on the prevalence of each dimension within our experience then more distinctive faces should be coded by the least prevalent dimensions (Hancock et al., 1996), and thus direction can detail distinctiveness. As faces are more densely clustered around the centre, those further from average should also appear more distinctive (Valentine et al., 2016). Although, paradoxically, very average faces are also atypical as few faces lie perfectly at the centre of face space on all possible dimensions (Burton & Vokey, 1998).

Norm-based coding is often described in terms of two opponent pools of neurons, with one pool firing maximally for one end of the dimension, the other maximally for the opposite end. The norm is coded by the overlap between the two opponent pools. This simple 'ratio model' was first outlined by Sutherland (1961), in the case of the motion aftereffect, and is described in detail by Susilo and colleagues (Susilo, McKone, & Edwards, 2010). In particular, Susilo and colleagues addressed whether the opponent codes have linear or non-linear response functions. To summarise, they suggested both are present.

In various, but not all, aspects of human face processing there is much behavioural evidence to support opponent, norm-based processing over exemplar coding using adaptation procedures. In adaptation procedures, the presence of an aftereffect infers some change to neural responses to the adapted stimulus, but only neurons sensitive to properties in that stimulus should be adapted. Adaptation in opponent coding should elicit an aftereffect in an opponent fashion. Opponent aftereffects have been observed for many aspects of faces such as race, gender and emotion (Skinner & Benton, 2012; Webster et al., 2004) and identity (Jiang et al., 2006, 2007; Leopold et al., 2001; Rhodes & Jeffery, 2006), however, as noted by Zhao et al (2011) opponent aftereffects can occur with multi-channel, exemplar coding.

Furthermore, aftereffects should selectively occur on the axis passing through the origin, corresponding to two opponent pools, and not affect stimuli orthogonal to that axis. This has been supported by evidence of larger identity aftereffects for stimuli which sit on the opposite sides of the origin (Leopold et al., 2001; Rhodes & Jeffery, 2006). The same has also been seen in the context of an expression space (R. Cook et al., 2011), where the dimensions of the space reflect changes in expression rather than in identity (e.g., Calder et al., 2001; R. Cook et al., 2011).

Smaller adaptation effects were also seen to stimuli not on the axis that passes through the origin, (e.g., Rhodes & Jeffery, 2006), which has been taken as evidence that adaptation effects the entirety of face space (Valentine et al., 2016) and could support exemplar coding. Yet, exemplar coding would predict equal aftereffects in any direction, orthogonal to the axis passing through the origin or not. The smaller effects may therefore be due to lower-level adaptation, or it could be that the stimuli did not sit exactly on an orthogonal plane, resulting in a slight adaptation effect.

Opponent coding would also predict that the aftereffect when viewing a neutral stimulus would increase with increasing adaptor strength. This has been observed, for instance with adaptation to facial expression (Skinner & Benton, 2012), and configural changes such as eye and mouth height (Robbins et al., 2007; Susilo, McKone, & Edwards, 2010). In contrast, a decline in the strength of the aftereffect would be predicted for exemplar-based coding once the adaptor strength is sufficiently distant from the norm, although this has been observed in a study on gender aftereffects (C. Zhao et al., 2011). McKone et al (2014) found interesting identity aftereffects. Identity adaptation aftereffects increased with increasing adaptor strength while in the range of natural variability, dropped slightly outside of this range, but then remained constant with no further decline. While the explanation for the small decrease in aftereffect is unclear, the lack of a further decrease is better explained by norm-based coding.

At the time when the study came out, an exemplar-based model, ‘Face-Space-R’ (Lewis, 2004) seemed to account best for various aspects of human face processing including distinctiveness, caricaturing and familiarity. However, the exemplar-based code does not explain many of the more recent findings. For instance, the decrease in discriminability between faces that are caricatured along the axis passing through the origin compared to changes in an orthogonal axis (Ross et al., 2010), highlighting the importance of direction relative to the norm over distance. Likewise, it does not explain the sustained or even increased adaptation aftereffect to stimuli distant from average, even past the realm of natural variability (McKone et al., 2014; Skinner & Benton, 2012; Susilo, McKone, & Edwards, 2010) nor the sustained fMRI response to caricatured faces (Carlin & Kriegeskorte, 2017; Loffler et al., 2005). The ramp-tuning of neural firing rates to stimuli of increasing distance from average in macaques (Chang & Tsao, 2017; Koyano et al., 2021; Leopold et al., 2006) is also not well explained by exemplar coding. These latter findings will be discussed in greater detail shortly.

As well as revealing evidence of norm-based or exemplar coding, adaptation studies also reveal both dissociations and overlap in the neural coding of different facial properties, suggesting either separate or shared representations. For example, aftereffects transfer across changes in identity but are weaker than same identity aftereffects (Fox & Barton, 2007; Skinner & Benton, 2012). This indicates both identity-independent and identity-dependent representations of expression, or a substantial cross-communication between identity and expression representations. Similarly, identity aftereffects can transfer across changes in view, but the amount of transfer varies with familiarity with the test face (Jiang et al., 2007).

1.2.2 PCA as a linear face space

Principal components analysis (PCA), a technique used commonly for dimensionality reduction, is often used for making computational models of face space. A more comprehensive description is provided in Chapter 2, but in summary it is a method for determining the most prevalent and correlated sources of variation in a dataset and describing them with orthogonal components. These components are hierarchically ordered based on the amount of variance they explain. As such, components which explain only a small amount of variance can be discarded. When applied to datasets of faces, the spaces returned are an example of a face space, with the average representation at the centre, and dimensions extending from that which encode combinations of features and/or textures.

While less suitable when learning millions of faces (Taigman et al., 2014), PCA might be suitable for representing 5000 familiar faces (Jenkins et al., 2018). PCA-based models have been used for facial expression recognition (Calder et al., 2001), identity processing (Andrews et al., 2023; Burton et al., 1999; Turk & Pentland, 1991), and have recently revealed that only a few components are crucial for identity recognition (Andrews et al., 2023). They have been used in models capturing within-identity variation (Aishwarya & Marcus, 2010; Beridze, 2021; Burton et al., 2011, 2016; Cowe, 2003; Shan et al., 2003) and can help explain aspects of human perception such as distinctiveness ratings and distinctions between hit rates and false alarms (Hancock et al., 1996), perceptual similarity ratings (Somai & Hancock, 2022) and the facial inversion effect (McCleery et al., 2008). Distinctive familiar faces are also recognised faster in PCA-based computational models than less distinctive faces, which are recognised slower by humans (Burton et al., 1999). PCA-based models also provide computational mechanisms behind expression processing in the amygdala and posterior superior temporal sulcus (Ahs et al., 2014; Said et al., 2010). As will be described later, there is also good evidence that neurons in macaques respond in comparable ways to PCA components, with neural responses corresponding to linear projections onto orthogonal axes of change (Chang et al., 2021; Chang & Tsao, 2017). Evidence for PCA-like representations have also been supported in lower-level vision (Hancock et al., 1992), with PCA of natural images extracting components similar in appearance to the bar and edge detectors used in early visual cortex (e.g. Burr et al., 1989).

PCA-based face spaces have also seen utility in forensic applications. For instance, EvoFIT (Frowd et al., 2004; Hancock, 2000) is a generative method for reconstructing portraits of suspects through evolving the weights on shape and texture components. The witness is presented with several images generated from a Gaussian distribution within the space, and rate how similar each is to the perpetrator. The system then alters the weights based on the similarity ratings, presenting a new array of images until the witness is satisfied. EvoFIT has been used extensively within the UK police force and overseas leading to many successful convictions (Frowd et al., 2019).

An advantage of how PCA holistically extracts correlated changes is its impressive capacity for reconstructing the whole face from only part of the image (Berisha et al., 2010; Turk & Pentland, 1991). From their computational model, Berisha et al (2010) found the eye and mouth regions the most informative for reconstructing the whole face and the remarkable reconstruction accuracies indicate these regions contain sufficient cues to the global configuration. The importance of these regions is consistent with behavioural evidence (Ince et al., 2016; Royer et al., 2016, 2018) and these diagnostic features have also been identified by Abudarham and colleagues (Abudarham et al., 2019; Abudarham & Yovel, 2016, 2020) for both unfamiliar face matching and familiar face recognition by humans and a Deep Neural Network (DNN). Moreover, the eyes have been found to be important in the N170 electrophysiological response during face detection (Ince et al., 2016), as measured using electroencephalography (EEG).

The recovery of missing areas of an image using the auto-associative memory of PCA (Valentin et al., 1994) has also recently been used to recover videos of the face from vocal tract MRI scans and vice versa (Scholes et al., 2020) and dynamic actions in one viewpoint from another (Beridze, 2021). These models take related but independent sources, such as the vocal tract scan and the video, and concatenate them prior to performing PCA. The analysis learns the correlated changes across both sources allowing one source to be recovered from the other. These models will be discussed more in Chapter 4, as the work in this chapter tries to build on the model created by Beridze (2021) to make it more biologically plausible.

The recovery of missing information using PCA might also explain why one of the core face-selective areas, the fusiform face area

(FFA, Kanwisher et al., 1997), shows both holistic and parts-based processing (Harris & Aguirre, 2010). The dimensions in PCA spaces, or face space more generally, can incorporate both global and more local changes. If the FFA holds a face space representation (as suggested by Carlin & Kriegeskorte, 2017; Loffler et al., 2005), then these dimensions could explain why the FFA showed adaptation effects consistent with both holistic processing and part-based processing when the top and bottom halves of the face were aligned and misaligned respectively. Forcing parts-based processing using misaligned face halves could prompt the recovery of the full face as seen in PCA models (Berisha et al., 2010; Turk & Pentland, 1991).

Rather than humans having a single face space representation, some research suggests different spaces are required for different groups. For instance, for different races (Armann et al., 2011; Jaquet et al., 2008) and different genders (Baudouin & Gallay, 2006; Griffin et al., 2011; Little et al., 2012). Griffin and colleagues (2011) made separate PCA spaces for male and female faces, and found that projecting an exemplar into the other gender’s PCA space allowed them to reconstruct an opposite-gender face showing ‘family resemblance’. However, they then found adaptation aftereffects transferred across genders, suggesting a shared neural population. It is possible that different ‘categories’ (e.g. genders) are encoded as clusters within this shared representation, each with a local, category-specific norm akin to suggestions of how identity is pooled (Abudarham et al., 2019), allowing both gender-specific and cross-gender aftereffects. However, Little et al

(2012) found no transfer of aftereffects across genders and Baudouin and Gallay (2006) showed mixed-gender morphs were rated as more distinctive than single-gender morphs, despite being closer to the global norm, suggesting they are processed relative to gender-specific norms rather than a global norm. Nevertheless, the use of local norms within a global space was somewhat the inspiration for the second PCA model created in Chapter 4 trying to achieve view invariance.

PCA-based spaces of facial motion are also gaining some traction (Beridze, 2021; Cowe, 2003; Nagle et al., 2013; Scholes et al., 2020; D. M. Watson & Johnston, 2022). Watson and Johnston (2022) for instance recently provided useful PCA-based methods for further investigating the spatiotemporal dynamics of motion using second-order PCAs. The first PCA analysis is performed across a series of frames for a given actor, with the frames containing both texture and shape deviations from a reference template. This forms an expression space (e.g., Calder et al., 2001; R. Cook et al., 2011) with the average facial expression at the origin. The frames for each repeat of a given sentence are then projected back into the space providing trajectories of loadings on the components. The loading trajectories across multiple repeats are then entered into a second PCA. This method allows one to manipulate and (anti)caricature the motion relative to the average trajectory for that sentence. Extracting regularities in facial motion (Furl et al., 2020; D. M. Watson & Johnston, 2022) within a face space representation might explain why perceptual grouping occurs for facial features that move synchronously (R. Cook et al., 2015; Johnston et al., 2021).

Overall, PCA has proved a powerful tool in creating computerised linear face space representations, however, one area that has seen little attention is that of view-invariance. Our work in Chapter 4 therefore aims to expand the PCA-based methods used by Beridze (2021) for learning and mapping facial motion across changes in viewpoint.

1.2.3 The role of experience on face space representations

The structure of human face space is thought to be optimised through experience (Short et al., 2011; Valentine et al., 2016; Webster & MacLeod, 2011), changing from infancy to adulthood (Hills et al., 2010; Short et al., 2011), but even in adulthood it remains flexible. For instance, Webster et al (2004) showed time spent in the US influenced the boundary in a morphed continuum where Japanese participants could distinguish Caucasian and Japanese faces.

Research suggests both similarities and differences in the face space representations in young children compared to those in adults. The distortions required to elicit certain adaptation aftereffects are larger in children (aged 5 and 8 years old) than in adults suggesting that their face space representation is less refined but is qualitatively similar (Short et al., 2011). Yet, unlike adults, Caucasian 5-year-olds had not fully developed a separate norm for Chinese faces (Short et al., 2011).

Furthermore, compared to faces with vertically aligned eyes, the aftereffect for faces with unnaturally misaligned eyes is just as strong in young children (ages 6-12 years) yet is much weaker in adults and adolescents (Hills et al., 2010).

The restructuring of human face space is thought to operate through adaptation. It has been suggested that the effects of adaptation reflect the updating of the average or norm through perceptual experience. This has been suggested in the contexts of identity (Rhodes & Jeffery, 2006), ethnicity, gender and expression (Webster et al., 2004), with adaptation leading to a shift in what appears average. The effect of adaptation can last for hours or days, even for familiar faces (Carbon & Ditye, 2012). This re-centring has also been supported by face-selective neurons in the macaque showing increased activity to novel faces with exposure, alongside concurrent changes in responses to familiar faces (Rolls et al., 1989).

A face space representation optimised through experience has the potential to explain why face-selective areas such as the occipital face area (OFA, Halgren et al., 1999; Puce et al., 1996) and the fusiform face area (FFA, Kanwisher et al., 1997) that are sensitive to lower-level visual information (Ramon et al., 2010; Weibert & Andrews, 2016; Yue et al., 2011), respond more to familiar than unfamiliar faces (Eger et al., 2005; Ewbank & Andrews, 2008) and why feedforward processing is faster for personally familiar faces than famous faces (Karimi-Rouzbahani et al., 2021). Familiarity does not appear to influence identity adaptation in the FFA (Weibert et al., 2016) suggesting the benefit of familiarity is not due to top-down influences of social attention but instead an optimised representation for familiar faces, even if sensitive to low-level image properties. That said, there is evidence that identity aftereffects only show when the adapting face is recognised (Laurence & Hole, 2012), indicating some reliance on top-down input and perhaps that facial identity aftereffects do not stem from the FFA.

The effect of familiarity on the transfer of aftereffects across views (Jiang et al., 2007) also suggests that face space representations for achieving view-invariance evolve over experience with individuals, not just faces as a general object class.

1.2.4 Neuroimaging evidence of face space representations

There is plenty of evidence to suggest a norm-based face space from adaptation in behavioural studies, and that PCA-style analyses might provide a route to making such representations, however, neuroimaging evidence for a face space representation and its cortical locus in humans in less clear.

The FFA, a patch of cortex sitting on the fusiform gyrus that preferentially responds to faces (Kanwisher et al., 1997), is thought to be a cortical locus of face space (Carlin & Kriegeskorte, 2017; Loffler et al., 2005), although not all evidence supports this (eg., Baseler et al., 2016; Davies-Thompson et al., 2013; Weibert et al., 2016).

Loffler et al (2005) firstly showed that the BOLD response (blood oxygenation level dependency) in the FFA increased as the distance increased between a synthetic face stimulus and the average stimulus. This followed an S-shaped tuning curve plateauing at a certain distance from average, but not decreasing, thus supporting norm-based coding over exemplar coding. In experiment 2 this increase was only significant when the origin (the average face) lay on the span of the plane being manipulated. I.e., modulating dimension 1 with all other dimensions set to 0, rather than varying dimension 1 with dimension 2 at a non-zero value (in this case, 9%). In this offset condition they showed adaptation to the same degree as presenting the same image repeatedly, showing the FFA adapted to the presence of component 2, further highlighting the importance of the origin. Experiment 3 showed a stronger release from adaptation to tangential changes in direction within the space compared to radial changes in the distance of the stimulus from the average. This suggests that different neurons in the FFA code different *directions* relative to the average rather than different *distances*, consistent with an opponent code.

A comparison of different computational models also showed that responses in the FFA to faces of varying distance from average were best explained by a norm-based account with S-shaped tuning curves rather than exemplar coding (Carlin & Kriegeskorte, 2017). That said, performance was similar in a model comprised of Gabor filters of varying sizes, orientations, positions, and spatial frequencies which did not explicitly code the direction or distance in face space.

In contrast to the aforementioned studies, others have shown no adaptation across multiple different frontal images of the same individuals within either the OFA or FFA compared to images of different people (Davies-Thompson et al., 2013; Weibert et al., 2016) suggesting against a face space representation. Theoretically, even with some variability in the components loaded onto caused by within-identity variation, there should be common components that all same-identity images loaded onto, causing neurons sensitive to that dimension to adapt, however this was not seen.

Baseler and colleagues (2016) also assessed changes in BOLD responses with identity adaptation, finding a stronger release from adaptation in OFA and FFA for changes influencing the shape and size of the face and its features non-linearly compared to linearly. This finding was interpreted as the OFA and FFA being involved in spatial alignment/normalisation prior to subsequent recognition.

Part of the problem with interpreting neural representations in humans comes from the limitations in methods. In humans, methods are generally limited to magneto/electroencephalography (M/EEG) and functional magnetic resonance imaging (fMRI). MEG and EEG have excellent temporal resolution but are more limited in spatial resolution. In contrast, fMRI has a better spatial resolution but is still generally limited to cubic voxels a few mm wide in each direction. It also measures changes in blood oxygenation as a (slow) proxy for neural activity. This limitation means that it is impossible to deduce exactly what each neuron is coding. Whilst invasive, single unit recordings (SUR) allow the responses of single neurons to be measured, and as such have revealed many interesting properties about face representations in macaques.

Recently, convincing evidence for face space representations has been found using SUR in the macaque face patch system (Chang et al., 2021; Chang & Tsao, 2017; Koyano et al., 2021). Chang and Tsao (2017) found neurons responding to combinations of shape and texture information, with linearly increasing firing rates to stimuli increasing in the neurons preferred dimension or ‘axis’ of change. Crucially for an opponent account of face space, neural responses were unperturbed by changes in orthogonal axes despite significant variations in appearance. This shows that rather than processing unique identities, the face processing system can describe faces as a linear projection onto orthogonal dimensions. Identity is then reflected in the relative weightings on the dimensions, consistent with prior suggestions from human studies (Ross et al., 2010).

Previous work by Leopold and colleagues (2006) also showed similar findings of ramp based tuning, in addition to evidence of V-shaped coding. In V-shaped coding neural responses were minimal at the average face and increased either side. This evidence was not consistent with the work by Chang and Tsao (2017) but more recent work has shed light on this discrepancy. Koyano and colleagues (2021) found that early neural responses (~100ms) showed ramp tuning as found by Chang and Tsao (2017), but later responses (~200ms) showed a down-regulation of more average faces. They interpreted this finding as reflecting lateral inhibition to the average, whereby the intermediate response of many neurons to the average stimulus (compared to a few to a non-average stimulus) causes dynamic suppression of the response to average stimuli. This down-regulation appeared as a ‘V’ superimposed on the ramp response function and could explain the V-shaped coding seen by Leopold and colleagues (2006).

This dynamic suppression of the average did not result from short-term adaptation and was present from the first trials in the block (Koyano et al., 2021). This shows that the average is naturally dynamically suppressed to maintain the representation of the norm-based face space.

The work by Chang and Tsao (2017) also suggests that the more posterior regions ML (middle lateral) and MF (middle fundus) are more dependent on shape features and the more anterior AL (anterior lateral) and AM (anterior medial) on non-shape features. Combined with previous reports of view-independence in AL and AM (Meyers et al., 2015) this suggests that surface information might either be more important than shape in forming a view-invariant representation, or, that by AM, shape is sufficiently aligned such that a single view-invariant representation of texture is left. In computation models, including textural information improves estimations of 3D shape from 2D images

(Feng et al., 2018), but at a cost of increased time. In macaques, it may be the case that surface cues also aid 3D representations for view-invariance, but the significantly increased computational costs cast doubt on whether surface cues are used in this way.

Although face space is often thought to be specific to faces, it appears that faces are more broadly represented within an object space, situated in the animate, 'stubby' quadrant of the identified 2D (animate/inanimate and stubby/spiky) object space (Bao et al., 2020), although many aspects of this space remain unknown at this current time (Bao et al., 2020).

1.2.5 Caricaturing in face space

Despite no prior natural experience with caricatures, such as those drawn by artists at the beach, we are remarkably good at recognising people, or at least seeing a resemblance, from caricatures. This demonstrates the perceptual system can, in this instance at least, extrapolate out from the range of natural variability. And, as already discussed, neural responses (Carlin & Kriegeskorte, 2017; Loffler et al., 2005) and adaptation aftereffects (McKone et al., 2014; Susilo, McKone, & Edwards, 2010) are sustained and even increased to caricatured faces (although see C. Zhao et al., 2011).

Caricaturing is essentially a continuum between being closer to (anti-caricaturing) or further away from (caricaturing) an average face relative to the veridical face. These variations can be made artificially, such as in the instance of artistic interpretations, whereby the difference between the face and the average face is artificially exaggerated to a large degree. In face space or PCA space, caricaturing can simply be performed by manipulating the loadings on the components. Caricaturing can also be performed on the spatiotemporal dynamics of facial motion (Furl et al., 2020; D. M. Watson & Johnston, 2022) rather than just structural changes underlying identity or expression.

Both exemplar and norm-based coding predict a behavioural advantage for caricatures, either from increased distances from other exemplars, or from increased distance from the average. This is supported through evidence that caricaturing line drawings and photographs enhances recognition (Kaufmann & Schweinberger, 2012; Mauro & Kubovy, 1992; Rhodes et al., 1987; Schulz et al., 2012), whilst anti-caricaturing (making the stimuli more average) leads to longer reaction times (Rhodes et al., 1987; Schulz et al., 2012). Caricaturing during learning also improves recognition of veridical faces (Rodríguez et al., 2009) suggesting it helps create representations for new faces. However, when matching unfamiliar faces, while modest caricaturing can be beneficial, caricaturing too much can be detrimental (McIntyre et al., 2013).

Caricaturing can also alter the stored representation of familiar faces. Adapting to extreme configural distortions shifted the perception of ‘veridical’ towards the distorted image even after only one presentation (Carbon & Leder, 2005). For clarity, Carbon and Leder explicitly state that their stimuli were not caricatured, as in their description caricaturing exaggerates distinctive features, however this thesis considers caricaturing to be any extrapolation from the norm, including configural changes.

The effect of caricaturing and distance from average has also been seen in neuroimaging studies, with increased neural responses to faces that are further from the average as measured using SUR in macaques (Leopold et al., 2006) and fMRI in human FFA (Loffler et al., 2005), even to profile silhouettes (Davidenko et al., 2012). Likewise, caricaturing exemplars form the norm also increases the EEG amplitude of the face-selective N170 and N250 event-related potentials (ERPs, Kaufmann & Schweinberger, 2012; Schulz et al., 2012). These findings are more consistent with norm-based coding rather than exemplar coding, which would predict a more uniform, if not a reduced response to faces that are further from the average. That said, other neural responses such as the P200, decreased with eccentricity (Schulz et al., 2012), indicating some neural processes encode averageness and typicality. This is perhaps unlikely to reflect dynamic suppression of responses to the average face (Koyano et al., 2021), which too would elicit smaller responses to non-average faces. Research suggests that an increased P200 amplitude indicates increased stimulus-directed attention (Bourisly & Shuaib, 2018; Picton & Hillyard, 1974) and that repetition suppression causes a decrease in the P200 (Freunberger et al., 2007), both of which are incongruent with active suppression of more average faces causing higher P200 ERPs.

Because caricatures are generally easily recognisable, and because of evidence of increasing neural response with increasing caricature level to relatively natural stimuli, we wanted to assess how the BOLD response would vary if face space was taken to the extreme. An experiment addressing this is detailed in Chapter 3, assessing the BOLD response to faces caricatured beyond the realm of natural variability.

1.2.6 Identity-general and identity-specific spaces

The initial idea of face space considered an ‘identity-general’ space, in which all identities are entered into a single space, but as has been discussed above, within-identity image variation is substantial. One source of within-identity variation is facial expressions, and explicit ‘expression spaces’ have been made to describe changes in expression (Calder et al., 2001; R. Cook et al., 2011). Given the large degree of variability across ages, genders, ethnicities and expressions, one of the highly debated topics of face space is how compartmentalised it is. In the context of identity and within-person variability, how far are identity and expression separated? And are all instances of a given person represented in an identity-general space or are there separate, identity-specific spaces for familiar individuals and if so, how?

Two such theoretical possibilities for dealing with within-identity variation in an identity-general space are discussed by Abudarham and colleagues (2019). The first is the Perceptual Single-Prototype Model, in which only the learned average representation of an identity sits within the space as a single point. This is consistent with the Average Face Model proposed by Jenkins and Burton (2011). The average representation becomes very stable once ~20 images have been learned, even if a few images of a different person are included. PCA models constructed from the averages across exemplar images also perform better in nearest neighbour recognition than those made using the images themselves (Burton et al., 2005).

Observers, including 3-month old infants (de Haan et al., 2001), naturally extract averages (Davis et al., 2021; Kramer et al., 2015; Neumann et al., 2018; Robson et al., 2018) even for dynamic faces (B. Chen & Zhou, 2018), thus the Perceptual Single-Prototype Model may be plausible. It is also consistent with some behavioural evidence that averages are recognised more easily than individual exemplar images (Bruce et al., 2002; Burton et al., 2005). Despite this evidence, averages are unlikely to be the only representation stored of a given individual. The average face does not reflect the best likeness of an individual (Balas et al., 2023; Ritchie et al., 2018) and other research shows that averages are recognised slower than exemplar images (Ritchie et al., 2018). Averaging also tells one nothing about how that identity’s appearance varies as a function of pose, expression or lighting (Burton et al., 2016). Finally, ensemble encoding occurs for inverted faces and other race faces (Davis et al., 2021) suggesting it is a general property of visual perception rather than being face specific. Therefore, even if an average is extracted and represented in an identity-general space, it is unlikely to be the only representation of a given individual.

The other model discussed by Abudarham and colleagues (2019) is the Conceptual Multiple Sub-Prototype Model in which images of a given identity are clustered together in face space, tied together using conceptual information. This is consistent with evidence that learning conceptual information improves face learning in humans (L. Schwartz & Yovel, 2016; Tanaka & Pierce, 2009; Yovel et al., 2012). It is also more commensurable with arguments highlighting the importance of variability (Burton, 2013; Burton et al., 2011) as well as evidence showing that variability is important for face learning (Murphy et al., 2015; Ritchie & Burton, 2017), more so than the average of an ensemble (Mondloch et al., 2023).

Glyn Cowe (2003) showed how different identities and the within-identity variation caused by expression could be incorporated within a single identity-general space based on opponent, norm-based coding. After aligning numerous identities to a standard shape, Cowe created a PCA-based representation using images of several identities taken over multiple expressions. The first dimensions coded appearance and shape changes important for identity. The next subset of dimensions coded rigid rotations of the head. Subsequent dimensions primarily coded non-rigid deformations caused by speech and were more individualistic and idiosyncratic.

More recently, the representations surrounding familiar faces have been described as ‘islands of expertise’ that sit within identity-general space (Hancock, 2021), but in a manner based on the Face-Space-R model of exemplar coding (Lewis, 2004) rather than norm-based coding. This theoretical model highlights the encoding of within-person variation of familiar individuals and is similar to the Conceptual Multiple Sub-Prototype Model (Abudarham et al., 2019) with familiar faces being clustered within an identity-general space. The ‘islands of expertise’ hypothesis (Hancock, 2021) also highlights why we experience seeing resemblance between novel and familiar people. Due to shared visual properties, the projections of the novel individual land within the vicinity of the known person’s ‘island’. The semantic knowledge of this resemblance then forms part of the code for the new person, with participants better at recognising novel individuals the more they resembled someone they already knew (Hancock, 2021). In contrast, novel faces who share no resemblance to known individuals are lost at sea, other than the approximate longitudes and latitudes given by the common dimensions.

In contrast to a single, identity-general space, identity-specific spaces are essentially the islands of expertise (Hancock, 2021), but where each island is coded by its own set of dimensions, not based on a set of common components. Each space captures the within-person variability during speech, changes in pose and other variable viewing conditions of a single familiar individual. They are essentially an expansion of person-specific expression spaces (Calder et al., 2001; R.

Cook et al., 2011). An early implementation of identity-specific spaces was by Shan and colleagues (2003) who made separate eigenface models (Turk & Pentland, 1991) for each individual, referring to each space as a face-specific subspace. Identity was determined by which space an image could be best reconstructed from. The same approach was also taken by Aishwarya and Marcus (2010) in their ‘multiple eigenface subspace’ model.

Both models (Aishwarya & Marcus, 2010; Shan et al., 2003) capture and explain idiosyncratic variation although both are limited in their methods. Other than aligning the images based on the eyes and mouth, shape was not normalised resulting in blurry, superimposed eigenfaces and reconstructions. In contrast, Cowe (2003), Nagle and colleagues (2013) and Burton and colleagues (2011, 2016) explicitly extracted and normalised shape information allowing the within-identity variation of rigid and non-rigid motion to be more clearly captured and expressed. In the models of shape parameters, Burton and colleagues (2011, 2016) consistently found that the first few dimensions reflect rigid head motion, such as changes in view and size, while subsequent dimensions encoded non-rigid motion such as expressions. Shape parameters are explained more in Chapter 2.

The separate spaces created by Burton and colleagues (2011, 2016) allow for the representation of idiosyncratic information that would be lost in the less prevalent components of an identity-general space. As proof of the extent of idiosyncratic variability, they projected novel images into spaces corresponding to the same or different identities. Reconstructions were worse when projected into different identity's spaces rather than their own, regardless of whether the texture or shape space was used. This highlighted how different information is captured across identities.

Whether identity-specific spaces are nested within an identity-general space has yet to be determined. It may even be possible for there to be parallel, non-nested spaces. Previous computer vision applications (Aishwarya & Marcus, 2010; Shan et al., 2003) have described the spaces as distinct, and recognition the process of projecting into all spaces and establishing which provides the best reconstructions. Their methods were based solely on computational goals whereas Burton and colleagues' (2016) were driven by understanding human behaviour, yet they too argued for separate, identity-specific spaces.

There are limitations of a system solely containing separate spaces. To recognise a face, it would either need to be projected into all spaces simultaneously, which would be quicker but have much higher short-term cognitive demands, or sequentially which would decrease short-term cognitive demands but dramatically increase the time taken. Moreover, having solely separate spaces also does not allow the transfer of aftereffects across identities (Fox & Barton, 2007; Skinner & Benton, 2012).

Theoretically, nesting identity-specific spaces within an identity-general space, i.e., as regions of the identity-general space with common axes (e.g., Hancock, 2021), would allow faster recognition as the common axes would direct the search to only the relevant ‘islands’. Rather than nesting the identity-specific spaces within the identity-general space, it may be possible to have two separate systems. Once candidate matches in the identity-general space have been established the input can be directed to the appropriate identity-specific spaces. The identity-general space may only need to represent a single template of each individual, as in the Perceptual Single Prototype Model (Abudarham et al., 2019), providing a cursory method for narrowing the search of which identity-specific spaces (if there are similar looking individuals) to project into. As these identity-specific representations evolve with experience, the identity’s template in the identity-general space can be updated.

1.2.7 Face space representations of facial motion

As highlighted above, facial motion is one of the major causes of within-person variability, through both rigid transformations from turning one’s head to non-rigid deformations around the eyes and mouth during speech. Systems for identity and gender recognition for instance need to be invariant to changes induced by motion. At the same time, speech and expression processing need to be sensitive to facial motion. How facial motion is represented in face space is therefore important for understanding how these two systems can function, and whether they require completely discrete systems or whether they overlap.

While face space applications are more limited in the context of facial motion, this area is gaining more interest. An early example is of course expression spaces (Calder et al., 2001) which explain the deviations of facial expressions from an average, neutral face. Cook and colleagues (2011) subsequently showed that expressions in these spaces are processed in a two-pool opponent manner.

In their PCA-based computational models, Cowe (2003) and Nagle and colleagues (2013) showed how expressions could be mapped across different identities using separate identity-specific spaces, again considering the expressions relative to an average, neutral position. Similarly, Beridze (2021) has begun to show how facial motion can be mapped across different viewpoints within a face space framework. These studies will be explained in further detail in Chapter 2 and Chapter 4.

In these expression spaces, dynamic sequences can also be considered as a trajectory through the space, with a timeseries of loadings onto the components. Not only can manipulations be made relative to the average, static stimulus, but they can also be made relative to the average dynamic trajectory for a given action. In a functional imaging study, Furl and colleagues (2020) varied and caricatured the displacement, speed and timing of dynamic expressions around their average trajectory. Caricatured movements were rated as more convincing, were more easily recognised, and increased the BOLD responses in the core face-selective regions compared to anti-caricatured movements (Furl et al., 2020). As caricaturing was performed relative to the average dynamic sequence for the expression, the caricatures could express more or less motion relative to the action’s norm. The increased response to caricatures therefore did not solely reflect the amount of motion or the magnitude of the displacement. This therefore poses questions about whether dynamic faces are processed within a single expression space, and if so whether they use a form of exemplar coding, or whether separate spaces for each known utterance or action are required. This further complicates the question raised by Dobs and colleagues (2018) as to how many dimensions are needed to fully encode the space of facial motion.

Although non-rigid facial motion, such as speech and expression, is separable from the rigid rotations that govern the observed viewpoint, and can be distinct from identity, it is important to consider how these factors interact and how they may be jointly coded within face space. As will be described in due course, motion provides cues to identity and aids view-invariant representations.

1.2.8 Viewpoint

Despite much research into view-invariance and face space separately, it is still not clear how view-invariance is achieved in such a representation, regardless of whether face space is nested or non-nested, or whether identity-general or identity-specific representations are present.

The identity adaptation supporting a norm-based face space partially transfers across views and the amount of transfer increases with familiarity (Jiang et al., 2007), indicating that different views share a neural response, and that familiarity helps establish this shared neural population. Nevertheless, it is not clear whether this supports a face space model containing a 3D representation or a few face space representations specialised for a few 2D prototypical views that are neuronally connected.

As the topic of achieving view-invariance is considerable, the next section will cover it in detail. It will describe the problem of view-invariance, current evidence for it in humans, macaques and computer vision systems, how facial motion effects view-invariance and what evidence there is currently for either 2D or 3D representations.

1.3 Achieving view-invariance

1.3.1 The problem of view-invariance

While there is not yet a sufficient explanation of how view-invariance is achieved in face space, either in the human brain or computational models, a substantial amount of research has been conducted on view-invariance. Here we aim to review findings from humans, macaques and computational models to highlight the challenges of view-invariance and gain insight into how it might be achieved and where in the brain.

Despite very impressive performance at familiar face processing, humans are much worse at processing unfamiliar faces (Jenkins et al., 2011; Ritchie et al., 2020), especially when variability between images is high (Sandford & Ritchie, 2021). This includes tasks that do not require explicit identity processing, such as orientation or gender discrimination tasks (Balas et al., 2007). The problem of unfamiliar face processing stems from the difficulties of dissociating within-identity variability (e.g., changes in pose, expression, lighting and occlusion) from between-identity variation (Jenkins et al., 2011). Pose can be especially detrimental due to the changes in the shape, illumination, and the occlusion of features such as of the far eye following rotation to a profile view.

When dealing with an unfamiliar face, representations are far from view-invariant (Bruce, 1982; Etchells et al., 2017; H. Hill et al., 1997; Longmore et al., 2008). When unfamiliar faces are learned from one view, the subsequent recognition shows a decreasing trend in performance as a function of angular distance (Etchells et al., 2017; Longmore et al., 2008), as does matching to a frontal view (Caharel et al., 2015). Learning multiple viewpoints improves recognition, but there is inconsistent evidence whether this aids recognition of novel intermediate views (Etchells et al., 2017; Longmore et al., 2008, 2015; L. Schwartz & Yovel, 2016), although this discrepancy possibly stems from the inclusion or exclusion of external features (Longmore et al., 2015).

Changes in viewpoint are encompassed by three axes of rotations; yaw, pitch, and roll, which can be caused for example by, shaking, nodding, and tilting the head respectively. The different rotations impact unfamiliar face processing to different degrees, with rotations in roll being less detrimental than changes in yaw, which in turn are less detrimental than changes in pitch (Favelle et al., 2007, 2011; Favelle & Palmisano, 2018; Van der Linde & Watson, 2010). Different rotations also tap into different processing mechanisms, with there being no interaction between changes in roll and changes in yaw (Van der Linde & Watson, 2010) including when fully inverted (180° roll, Favelle et al., 2017). The benefit for yaw rotations over pitch possibly stems from to ability to use symmetry for some changes in yaw (Favelle et al., 2017).

With increasing familiarity, view-invariance is developed. Recognition over views is much better for familiar faces than unfamiliar faces (Bruce, 1982; H. Hill et al., 1997) and figural aftereffects transfer more strongly over viewpoint with increased familiarity (Jiang et al., 2007). However, how view-invariance is achieved is yet to be fully established. Theoretically, learning within-identity variation could provide information on metrics that are invariant to changes in the face, such as distances between certain features (Burton et al., 2015). However, stretching images of familiar faces has no effect on recognition (Burton et al., 2015) and participants are no better at stretching or contracting familiar faces back to the right aspect ratio than unfamiliar faces (Sandford & Burton, 2014), arguing against learning specific invariant metrics for familiar faces. Furthermore, inter-feature distances inevitably change with viewing angle. The metrics might be processed in 3D rather than 2D, but even then, the results of Sandford and Burton and Burton and colleagues suggest these metrics are not reliable.

How view-invariance is developed, how it interacts with identity and other factors such as facial motion, and how it fits with a face space framework has yet to be fully discovered.

1.3.2 The role of motion

Early theoretical models of face processing originally separated dynamic aspects of faces from invariant ones, such as expression versus identity (e.g., Haxby et al., 2000) and these processes seem to be separated in the cortex. However, there is a body of evidence showing that facial dynamics impact the perception of other factors, such as identity processing and view-invariance. Elaborating on how facial motion interacts with identity processing and face space is therefore of much importance.

In general, facial motion in humans is thought to be primarily processed in the posterior, mid and anterior clusters of the STS (Haxby et al., 2000; Pitcher et al., 2011, 2014; Polosecki et al., 2013; H. Zhang et al., 2020) and in the right inferior frontal gyrus (Nikel et al., 2022; Pitcher et al., 2011) whereas identity processing is thought to reside more in ventral regions such as the FFA (Haxby et al., 2000).

More anterior regions of STS become more selective for *facial* motion over either body or object motion (Pitcher et al., 2011; H. Zhang et al., 2020) and to non-rigid deformations over rigid motion (H. Zhang et al., 2020). The aSTS may therefore play a crucial role in expression and speech perception. That said, there is still evidence that rigid motion (leftward versus rightward changes in yaw) can be decoded in the aSTS (Carlin et al., 2012).

The representation of dynamic facial stimuli in the STS is also consistent with evidence showing that an individual with acquired prosopagnosia was selectively impaired on recognising static but not dynamic facial expressions (Richoz et al., 2015). Her lesion covered the right inferior occipital gyrus (the location of the OFA) but spared the STS. The STS has also recently been suggested to form part of the third, previously undefined visual pathway solely for processing dynamic social signals from the face and body (Pitcher & Ungerleider, 2021).

While obviously important for speech and expressions, facial motion also interacts with other factors such as identity. Evidence shows motion: is sufficient for identity and gender judgements (Girges et al., 2015; H. Hill & Johnston, 2001), aids the learning and recognition of unfamiliar faces (Lander & Davies, 2007; see review by Lander & Pitcher, 2017) and can improve recognition when images are visually degraded (Knight & Johnston, 1997; Lander & Bruce, 2000; also see O'Toole et al., 2002) or when form is 'unreliable' (Dobs et al., 2017).

Even in the presence of identity-specific form, motion of another can bias perception of identity (Knappmeyer et al., 2003). Although, not all results show a consistent advantage for motion, particularly if the expression changes between learning and test (Christie & Bruce, 1998).

There are currently three main, complementary theories of how motion improves recognition and some theories have even suggested it aids view-invariance (Lander & Butcher, 2015; Lander & Pitcher, 2017; O'Toole et al., 2002). They suggest that motion provides supplementary information about identity through idiosyncratic movements (Supplementary Information Hypothesis), that it enhances the structural representation of the face (Representation Enhancement Hypothesis) or that it provides no useful information *per se* but instead directs social attention to the face which in turn benefits identity processing (Social Signals Hypothesis). The second of these theories could contribute to achieving view-invariance through enhancing flexible structural representations. As noted by Furl and colleagues (2017), physical structure is important for identity, but it also constrains the physical range of facial motion and thus the perception of structure from motion.

In support of the Supplementary Information Hypothesis, Lander and Davies (2007) only found a benefit of learning faces in motion when they were presented in motion both at learning and at test. They therefore argued that their results demonstrate recognition through learning "characteristic motion signatures", an idea initially proposed by Knight and Johnston (1997), rather than aiding recognition from structure. Supporting this, Lander and Chuang (2005) only found a benefit for facial motion for faces whose motion was rated as being distinctive.

While Lander and Davies (2007) found no behavioural benefit of learning motion on subsequently static faces, other results have found such an advantage in old/new recognition (Butcher et al., 2011; Pike et al., 1997), matching without memory (Thornton & Kourtzi, 2002) and in a visual search paradigm (Pilz et al., 2006). These results are consistent with the Representation Enhancement Hypothesis, although they can also be explained by the Social Signals Hypothesis.

One route to achieving view-invariance through rigid facial motion is the close temporal proximity between different views (Wallis, 1996, 2002; Wallis & Bülthoff, 2001). Wallis (2002) showed that when images of different identities were presented in an apparent smooth transition across views (one identity per view), it was then harder to recognise these individuals as different identities. This relied on the smooth temporal association between plausible neighbouring viewpoints; the effect was abolished when the viewpoints were presented in a random order. Recognition of objects is also easier when the presented view continues the previously observed sequence of rotation, compared to a view rotated in the opposite direction from the starting view (Vuong & Tarr, n.d.). For faces and objects we therefore build an invariant representation by learning instances closely related in time, and prime the representation of the view we predict will be seen next.

There is mixed behavioural evidence over which type of motion benefits identity recognition the most. Hill and Johnston (2001) observed an advantage of non-rigid motion for gender judgements but an advantage for rigid motion for identity, whereas Lander and Chuang (2005) observed an advantage for non-rigid motion on identity judgements. Separating non-rigid motion into semantic groups, Lander and Chuang (2005) saw comparable advantages for both communicative and expressive non-rigid motion, yet Dobs et al (2016) found an advantage for communicative action, regardless of whether expressive actions were within or outside of social contexts.

The coherence and meaning of motion is also important. The dependence on the natural ordering and temporal dynamics of frames (Dobs et al., 2014; Lander & Bruce, 2000) and the benefit of rigid motion compared to either single or multiple static images (Pike et al., 1997 but see Mileva & Burton, 2019) suggests the role of motion transcends the benefit of simply having more visuospatial information in a video. The use of non-rigid motion increases with familiarity and idiosyncratic motion is thought to be represented in the posterior superior temporal sulcus (O’Toole et al., 2002), which has been shown to be sensitive to identity (Ramon et al., 2010; Sliwinska et al., 2019). The integration of motion and identity, perhaps through idiosyncratic actions, may occur in the pSTS, although, aSTS which shows a higher level of selectivity to non-rigid deformations is likely another candidate (H. Zhang et al., 2020).

There is little evidence as to whether non-rigid or rigid motion aids view-invariance. In their second experiment on matching facial motion across views, Watson et al (2005) found that extrapolating across viewpoints was as effective as interpolating viewpoints from non-rigid motion. Sample stimuli were presented at either 15° and 45° or 45° and 75°, with test stimuli then presented at 30° or 60°. When rigid motion was added, participants were worse at extrapolating from more frontal views to more profile views. This indicates that non-rigid motion is processed in a more view-invariant manner, whereas rigid motion is more view-dependent. Non-rigid motion of course is independent of view, while an animation performing rigid motion may be confounded by essentially a wider change in rigid motion through changing the viewpoint. It remains to be seen how rigid and non-rigid motion aid view-invariance for identity recognition.

Recent research has further contested the separation of variable and invariant aspects, leading to the suggestion of the Integrated Representation of Identity and Expression Hypothesis (E. Schwartz, O'Nell, et al., 2023). Schwartz and colleagues showed that deep convolutional neural networks (DCNNs) trained to recognise either facial expressions or identity were able to distinguish the other category, despite not being trained on it (E. Schwartz, Alreja, et al., 2023; E. Schwartz, O'Nell, et al., 2023). Performance increased in higher levels of both DCNNs even though the features used became more segregated, suggesting they separated but retained information about the untrained variation. Furthermore, the DCNNs did not differ in how well their responses to stimuli (varying in identity and expression) correlated with neural responses within lateral and ventral regions. The DCNN trained to recognise identity was slightly more correlated with both regions than the DCNN trained on expressions.

The retention of task irrelevant information has also been observed in other models trained on identity recognition, for example ResNet101 retained both expression and viewpoint information (Colón et al., 2021) and VGG-Face retained "hallmarks of human expression recognition" (L. Zhou et al., 2022). Overall, this work shows that ventral and lateral regions of the brain may be more functionally similar than initially predicted when processing dynamic and invariant information. Rhodes et al (2015) also showed that identity and expression information can be coded by common dimensions in face space, supporting the more identity-specific dynamic components reported by Cowe (2003) and further highlighting that identity and expression may not be as easily separated as prior work suggests.

While the response in the ventral cortical regions, such as the OFA and FFA to dynamic faces was not consistently decreased following theta burst stimulation to the rpSTS (Pitcher et al., 2014, 2017), the resting state connectivity between these regions was reduced (Handwerker et al., 2020), suggesting that these regions might also functionally interact. Moreover, attention to identity has been found to increase decoding of identity both in ventral and dorsal regions (Dobs, Schultz, et al., 2018) and sensitivity to expression has often been observed in the FFA (see review by Bernstein & Yovel, 2015), further highlighting overlap in function.

Although motion improves face matching in patients with congenital prosopagnosia (face blindness) it does not improve recognition relying on memory, suggesting that longer-term representations of motion are harder to encode than short-term representations (Longmore & Tree, 2013).

To summarise, non-rigid facial actions are crucial for visual speech and expression processing, yet also contribute to identity processing and possibly achieving view-invariance. Moreover, recent research suggests that areas thought to process either identity *or* expression may represent both. Thus, it seems fruitful to understand how identity and motion are represented together in face space.

1.3.3 View-invariance in the macaque face processing system

The macaque visual system has provided incredible insight into how the human brain might function due to structural and functional similarities (Pinsk et al., 2009; Tsao et al., 2008), and through the ability to record directly from the cortical surface using SUR. Face-selective cells in macaques are predominantly clustered into a network of discrete patches running down the upper and lower banks and fundus of the posterior, middle, and anterior STS (Baylis et al., 1987; Freiwald & Tsao, 2010; Perrett et al., 1988; Pinsk et al., 2005, 2009; Tsao, 2006; Tsao et al., 2003).

Early research into view-(in)dependence revealed neurons preferentially tuned to different prototypical views, including frontal (0°), profile (±90°), back, upwards and downwards tilted views, with fewer specifically tuned to ¾ views (e.g., 45°) and even fewer to intermediate views (22.5° Perrett et al., 1991). View-selective neurons are seen in partially overlapping clusters (Perrett et al., 1988) with neighbouring views occupying neighbouring regions of cortex (G. Wang et al., 1996, 1998). Some other neurons show invariance to mirror views (Desimone et al., 1984; Perrett et al., 1991) and others to all views (Perrett et al., 1988, 1991). It was argued these cells are hierarchically organised from view-dependence to view-independence (Gross & Sergent, 1992). Even view-selective neurons however show large tuning widths, responding to a broad range of views (Perrett et al., 1991).

Understanding the preference for prototypical views is important for uncovering how faces are processed and has practical implications for advancing face processing software. Given that faces are observed from all viewpoints it is likely that this prototypical preference paired with broad tuning profiles is computationally sufficient (or perhaps optimal) for constructing a view-invariant representation. As will be discussed in the following section and in Chapter 6, there is currently limited evidence for which views are preferentially processed in humans.

Returning to view-invariance in macaques, recent evidence using multi-voxel pattern analysis suggests more posterior areas, MF and ML ('middle fundus' and 'middle lateral', Pinsk et al., 2009; Tsao et al., 2008) encode identity with respect to view. In contrast, more anterior regions AL and AM ('anterior lateral' and 'anterior medial', Pinsk et al., 2009; Tsao et al., 2008) encode identity in a more view-independent manner (Meyers et al., 2015), with AL generalising across mirror views (Freiwald & Tsao, 2010).

Adaptation techniques however have also indicated view-invariance in both area AF ('anterior fundus') and MF (Taubert et al., 2020). View-invariance in AF is not surprising given the invariance in AM and AL, yet invariance in MF is at odds with prior research (Meyers et al., 2015). This discrepancy may be due to the task, with view-dependence during identity processing (Meyers et al., 2015) but view-invariance during expression processing (Taubert et al., 2020). This would be compatible with evidence in humans that non-rigid facial motion is processed in a view-invariant manner (T. Watson et al., 2005).

Surprisingly, recent research suggests neurons in AM and AF are sensitive to local features such as the mouth and eyes rather than the global configuration (Waidmann et al., 2022). As the authors discussed, this was surprising given that the general assumption that neurons in more anterior areas are more holistic, but it is consistent with suggestions that configural processing is less important than once thought (Burton et al., 2015; Sandford & Burton, 2014). It also raises questions as to whether preferences for certain axes in the study by Chang and Tsao (2017) can be explained by local featural rather than holistic changes within dimensions. Finally, given the view-invariance in AM and AF, it suggests that local features might provide the best route to view-invariance.

1.3.4 View-invariance in the human face processing system

Like the macaque face processing system, the human face processing system comprises a system of interconnected regions. These include, but are not limited to, the occipital face area (OFA, Halgren et al., 1999; Puce et al., 1996) sitting on the lateral occipital cortex, the fusiform face area (FFA, Kanwisher et al., 1997) sitting on the fusiform gyrus, the posterior superior temporal sulcus (pSTS, Morris et al., 1996, 1998), the anterior temporal lobe (ATL, Rajimehr et al., 2009) and the inferior frontal gyrus (IFG, Vignal et al., 2000). Although, not all face-selective neurons are housed within these clusters (Schrouff et al., 2020).

As in macaques, the human visual system moves from a view-dependent to a view-independent representation along the posterior-anterior axis of the temporal lobe. There is generally evidence that the OFA, FFA and pSTS process faces in a more view-dependent manner (Ewbank & Andrews, 2008; Fang et al., 2007; Guntupalli et al., 2017; Natu et al., 2010; Pourtois et al., 2005). Yet, some view-generalisation has been observed in the OFA (Anzellotti et al., 2014), FFA (Anzellotti et al., 2014; Ewbank & Andrews, 2008; Guntupalli et al., 2017; Ramírez et al., 2014) and the pSTS (Natu et al., 2010). An intermediate level of view-invariance has been observed in the right ATL (Anzellotti et al., 2014; Guntupalli et al., 2017) and full view-invariance within the left middle temporal and left inferior frontal cortex (Pourtois et al., 2005) and in the right inferior frontal face area within the right IFG (Guntupalli et al., 2017). A moderate amount of view-invariance (at least up to +/- 20^{o} in yaw) during gaze processing has also been found in the anterior STS (Carlin et al., 2011).

Ramírez (2018), however, noted that the evidence of view-invariant identity discrimination in the FFA and ATL (Guntupalli et al., 2017) could be confounded by gender, as the identities contained two male and two female faces. Thus, it may be gender that is processed across views, not identity. In the 2018 paper, Ramírez argues that it is unknown whether most regions of the human face processing network represent faces in a view-invariant manner.

The intermediate view-invariance observed in the FFA might reflect the use of mirror-symmetry. There is some evidence to support this (Flack et al., 2019; Rogers & Andrews, 2022), however Ramírez and colleagues (2014) found that the FFA is sensitive to angular distance but not mirror-symmetry, suggesting view-dependence. In contrast, Anzellotti and colleagues (2014) found that identity can be classified in both the FFA and the OFA across views even when mirror-views are excluded, suggesting view-invariance that transcends mirror-symmetry. Yet, Weibert and Andrews (2016) found that the response pattern in the FFA was best explained by viewpoint than by identity, suggesting that the representation is not view-invariant. Furthermore, the response patterns could be explained by modulations in low-level image properties between changes in view and identity (Weibert & Andrews, 2016).

Evidence of view-invariance in the fusiform gyrus, possibly within the FFA has also been seen with EEG and the results indicate view-invariance for new faces is developed rapidly (Zimmermann & Eimer, 2013). The N250r component, thought to emerge from the fusiform gyrus (Schweinberger et al., 2002), occurs for repeated presentations of a single identity and reportedly reflects matches being made between a new image and face memory, therefore reflecting recognition of repeated identities (Schweinberger & Burton, 2003). Zimmermann and Eimer (2013) found this N250r deflection extended across changes in view after only a relatively short amount of training (~25 x 200ms exposures to each viewpoint) to an unfamiliar identity. This suggests view-invariant representations are constructed rapidly and that view-invariance is developed in the fusiform gyrus. Zimmermann and Eimer only tested with frontal and ~35^{o} views, both of which were present during learning. Caharel and colleagues (2015) found present, but generally weaker N250r deflections even over large changes in view, suggesting a larger degree of view-invariance, but not a fully view-invariant representation.

Although the FFA is typically referred to as one region, there is evidence from both recordings from electrodes placed on the cortical surface (Schrouff et al., 2020) and from fMRI (Weiner et al., 2014) that there are two patches, referred to as the posterior fusiform (pFUS/FFA-1) and middle fusiform (mFUS/FFA-2). Grill-Spector and colleagues (2017) suggest that more research should investigate the functional differences in viewpoint processing between these regions.

Problematically, detection of view-invariance varies with analysis, for example view-invariance was found in the FFA and ATL using multi-voxel pattern analyses in pre-defined regions of interest (ROIs) but not in unconstrained searchlight analyses (Guntupalli et al., 2017). Moreover, Ramírez (2018) showed that when angular distance is used rather than Euclidean distance to compare response patterns, the FFA does not show sensitivity to mirror views and instead shows sensitivity to angular distance.

Over human and macaque face processing systems, it is clear that there are responses to faces that are view-invariant, but, how and where this view-invariance is achieved is still not fully understood. One way to explore the potential options of how view-invariance might be achieved is looking at computational models and artificial neural networks.

1.3.5 View-invariance in computational models of face processing

In recent years performance of automatic face recognition systems has improved so much that they now reach, and exceed, human level performance (Hancock et al., 2020; also see review by Phillips et al., 2018). They use a variety of methods for achieving view-invariance (see Ding & Tao, 2016), and while achieving view-invariance has come a long way over the last few decades, it still remains a stumbling block (Ding & Tao, 2016). Here, a few of these methods are introduced.

One overarching method that the work in Chapter 4 draws from is to transform faces to a canonical view, often frontal, prior to analysis. This is often referred to as ‘frontalisation.’ Initial attempts tried image-based warping, however this was unsuccessful over large rotations, resulting in unnatural distortions (Berg & Belhumeur, 2012; Chai et al., 2003). Rather than using landmarks or image segments, Beymer and Poggio (1995) transformed a novel face image to a ‘virtual view’ using optic flow fields. Firstly, optic flow fields were calculated to estimate the pixelwise displacements across viewpoints of a set of ‘prototype’ faces. These learnt displacement fields could then be applied to a novel image. Again, however, unnatural distortions were present in the reconstructions rotated ±30^{o} in yaw and ±20^{o} in pitch.

More recent methods include a cascade regression procedure (Y. Wang et al., 2022) using a series of regressions to predict the shape of the frontal view from an arbitrary pose, rather than a single step. The output is combined with that of an active appearance model (AAM) trained on frontal views, producing an accurate reconstruction and minimising the effect of self-occlusion from profile poses. Together they allow the capture and translation of facial expressions across views. While using a series of regressions, the model still translates directly from any view to frontal, whereas an alternative method would be to cascade around neighbouring viewpoints. The realisation of occluded features across neighbouring viewpoints might allow for a better transformation across views.

Other options use a 3D model, for instance to aid with viewpoint estimation and landmarking for improving the efficacy of 2D warping (Taigman et al., 2014) or for rotating the face on a 3D model. The shape and/or texture can be rendered into 3D, such as by rendering the texture onto a 3D model morphed to the 2D image shape (A. Asthana et al., 2011), or morphing the 2D image to the 3D model shape (Hassner et al., 2015). These models can incorporate occlusion detection and extort mirror-symmetry to ‘fill’ occluded regions, although the model by Asthana and colleagues was still only effective up to ~45° changes in yaw and 30° changes in pitch. Researchers have noted however that 3D representations come with additional storage demands and inferring the 3D structure from a 2D image is an ill-posed problem (Zhu et al., 2013), thus a 2D-based system seems advantageous. If trying to recapitulate human processing, humans are only exposed to one view at a time, and so a 3D representation might be more challenging to build than from the 3D laser scans often used.

Such 3D models include for example the 3D Morphable Model (3DMM Blanz & Vetter, 1999) which comprises shape and texture spaces established from PCA analyses of 200 laser scans while actors produced different facial expressions. Blanz et al (2002) showed that novels views could be reconstructed from a single image, by first determining the linear combination of texture and shape components that best reconstruct the image. Without frontalising, identity can be determined by finding which identity in a gallery set has the closest matching shape and texture features to the test stimulus (Blanz et al., 2002).

Rather than calculating the parameters of a 3D morphable model, Jackson and colleagues (2017) showed that it was possible to learn 3D volumetric structure by training a CNN with 2D images and 3D laser scans. The model learned through regression how to predict the 3D structure from the 2D image and could thus reconstruct the 3D shape of images under varying poses and expressions. However, the model struggled to estimate 3D structure from profile images.

The fact that Jackson and colleagues' (2017) model could capture expressions is beneficial, as other methods often lose the dynamic properties. For example, Zhu and colleagues (2014b) reduced the within-person variation caused by different poses through frontalisation, but in doing so lost much of the non-rigid deformations. This would hinder view-invariant expression and speech perception and possibly even the perception of identity from motion (e.g., H. Hill &

Johnston, 2001). However, Jackson et al (2017) were only able to capture expressions well when there was a tight spatial correspondence between the images used for training and the 3D volumes.

Instead of trying to warp images across views, an alternative view-transformation method is to learn, and then map across, separate view-specific representations (Beridze, 2021; Lan et al., 2012). In their first model, Beridze made a PCA space for each viewpoint. Vectors describing the difference between individual frames and the average image for that viewpoint were calculated, and then projected into the other views' spaces. Beridze could reconstruct motion across small changes in view well, but not larger changes. In the first of our models, we attempt to map across larger changes in view through a cascading procedure of projecting and reconstructing frames across neighbouring views.

In a subsequent model, Beridze (2021) concatenated videos together, captured simultaneously from multiple viewpoints. PCA was performed on these multi-view vectors resulting in components containing information about all viewpoints, nested in distinct portions or 'slots' of the vectors. By projecting a single view onto the multi-view components Beridze was able to reconstruct non-rigid deformations across the remaining views. The model focused on identity-specific spaces for mapping facial motion across views, but it could be expanded to incorporate multiple identities using methods outlined by Cowe (2003). Our work in Chapter 4 expands on Beridze's model to better mimic human processing by removing the necessity for the model to ‘see’ multiple views at once.

Rather than projecting vectors directly across different views or creating multi-view vectors, Lan and colleagues (2012) instead used regression to learn how features in one viewpoint could predict the features in another, and so could reconstruct actions across wide changes in viewpoint. Their work solely focused on the mouth due to their interest being in automatic lip-reading software. Their method again required simultaneously recorded videos across wide viewing angles in order to establish the mapping from, say, profile to frontal.

Cross-view transformations are common, but are not the only method for dealing with different viewpoints. As an example, other models instead opted for separate, view-specific representations without mapping across views, aligning images to one of a few prototypical 2D view templates (An et al., 2019; Pentland et al., 1994). This avoids problems with warping and frontalising, but requires a suitable set of template images, a suitable number of template views, and sometimes separate processing systems for each view that require distinct features. This can be more computationally demanding than having a single view that all inputs are transformed too. In the context of automatic lip reading software, Lan and colleagues (2012) suggest it is less computationally demanding to rotate the lips to a canonical view rather than having separate view-specific systems.

An alternative approach is to identify ‘identity preserving’ features that minimise pose-dependency. These models use pose-robust feature extraction to, for example, sample smaller regions around the eyes, mouth and nose to use as features (D. Chen et al., 2013). As they are cropped to a small area they are less effected by viewpoint than more holistic methods such as the eigenface method (Sirovich & Kirby, 1987; Turk & Pentland, 1991). Chen and colleagues (2013) showed that recognition increased with increasing number of landmarks used and with multiple scales of sampling around said landmarks. Cao et al (2010) showed that recognition can be improved by weighting the features based on which features contribute the most to the given pose. The recent discovery that the more view-invariant (Meyers et al., 2015) AM and AF neurons in macaques are more sensitive to local features than holistic representations (Waidmann et al., 2022) provides support for these models.

Extraction of these features has also been effectively combined with a reconstruction layer that can reconstruct any face in a canonical pose, such as frontal (Zhu et al., 2013), or a parallel representation of viewpoint that allows reconstruction under any view (Zhu et al., 2014a). Furthermore, DR-GAN (Tran et al., 2017) combines the use of both identity-preserving features and viewpoint transformations, and a d-CNN (X. Yin & Liu, 2018) combines view-invariant representations with view-specific representations, with both studies improving recognition over using a single method.

Interestingly, pose-invariance in neural networks can also develop spontaneously without explicit instruction about 3D shape. A modified version (H. Lee et al., 2020) of a deep artificial neural network (AlexNet, Krizhevsky et al., 2017) was trained on object categorisation whilst minimising a proxy for the spatial distance between coactivated units, simulating reduced wiring distance between coactivated neurons. Despite not being provided with explicit information about 3D structure this model showed an increase in view-invariance in higher layers. It also shared other properties of inferotemporal cortex such as clusters of face units. Abuhdarham et al (2021) revealed a similar hierarchical transition from view-dependence to view-invariance in a different DCNN trained on faces, including revealing intermediate mirror-symmetric responses that prior models such as VGG-16 have failed to show (Yildirim et al., 2020).

Although DCNNs achieve remarkable performance and show striking similarities to the neural representations in the brain (Abudarham et al., 2021; H. Lee et al., 2020; Yamins et al., 2014; Yildirim et al., 2020), they too are self-learning computers, learning to classify the visual input based on experience and, crucially, based on limited tasks such as solely on identity recognition. Thus, it can sometimes be difficult to take these models and learn something about the computations in the brain, although there are instances when these models lead to interesting discoveries (e.g., Bao et al., 2020).

Furthermore, these networks make non-human like errors, such as mistaking identities of different genders and races (Hancock et al., 2020), casting doubt on their applicability to human face processing. Likewise, they can become too invariant to certain properties. VGG-Face (Parkhi et al., 2015) for instance is overly invariant to illumination compared to face-selective neurons in macaques (Chang et al., 2021). In fact, recent work suggests a simple active appearance model can better predict neural responses than many DNNs (Chang et al., 2021), with the exception being a model training on general object classification.

Overall, a common problem for computational models is translating over a large change in pose or expression, and many lose information about expression in favour of a more rigorous alignment for recognition. Given that facial motion aids recognition and view-invariance (Furl et al., 2020; H. Hill & Johnston, 2001; Knappmeyer et al., 2003; Lander & Bruce, 2000), we continue to build on the multiple appearance model described by Beridze (2021) and colleagues in Chapter 4 to allow the mapping of non-rigid deformations across changes in view. Our work develops this model to improve biological plausibility and provide a potential mechanism by which the brain might achieve view-invariance.

1.3.6 2D versus 3D representation

In 2000, Hancock and colleagues highlighted that advancing computer technologies would allow the importance of 3D information to be explored, and posed the unanswered question of “How important are motion and 3-D information in face identification?” (Hancock et al., 2000). It is evident that view-invariance in humans is developed over time (Gliga & Dehaene-Lambertz, 2007; Ichikawa et al., 2019), yet, over 20 years later it is still unknown whether view-invariance is achieved with 2D or 3D representations. In computer systems, both 2D and 3D methods have been adopted, but it is not clear which the brain utilises.

Hill and colleagues (1997) argued for the necessity of a 3D representation, however evidence from novel object recognition favours 2D (H. H. Bülthoff & Edelman, 1992), with recognition being worse for novel views outside the range of previously learned views. Similar results have also been found with faces (W. Chen & Liu, 2009; H. Hill et al., 1997; Y. Lee et al., 2006; Liu et al., 2009; Schwaninger et al., 2007) with the ability to recognise a novel face diminishing with increasing angle between learned and test views, supporting a 2D interpolation account, at least for unfamiliar faces.

This effect is particularly prominent for changes in pitch after learning changes in yaw (Schwaninger et al., 2007), which although perhaps less encountered frequently than changes in yaw, should still be recognisable with a 3D representation. Likewise, although recognition across changes in yaw was improved by the rigid motion of either passively viewing, or to a greater extent, actively exploring avatars in a virtual reality (VR) environment, recognition across changes in pitch was still poor (I. Bülthoff et al., 2019). These faces were unfamiliar to the participants however, so it may take time to learn the 3D representation, either as a unique 3D model or as a deviation relative to a single template.

Further evidence for a 2D system comes from behavioural evidence for categorical representations of view, with 0° and 6.7° and then 13.3° and 20° being grouped together (Y. Lee et al., 2006) suggesting grouped, view-tuned representations that are not well explained by a 3D representation.

The fact that the brain can be tricked into thinking different identities are one when temporally associated across smooth transitions in view (Wallis, 2002; Wallis & Bülthoff, 2001) further supports a 2D interpolation account of face processing, at least for unfamiliar faces. If based on a 3D representation, then it should not be possible to group different identities together that have different facial structures.

Further behavioural evidence against a 3D representation comes from adaptation aftereffects, showing that opponent viewpoint aftereffects for discriminating which direction a frontal view is facing peak at adaptors 20-30° from frontal (J. Chen et al., 2010). If faces are represented solely with a 3D model, then the aftereffect should not differ across viewpoint.

In contrast, other evidence suggests a 3D representation. For instance, Jiang and colleagues (2009) found that learning from two viewpoints resulted in a more illumination-invariant representation than learning from one view. The interpretation was that learning from multiple views increases the ability to represent 3D structure, independent of illumination. This increased illumination invariance is not well accounted for by 2D interpolation.

Zhao and colleagues (2016) reported some additional evidence in favour of 3D processing, showing that in the composite face effect (A. W. Young et al., 1987), removing 3D cues by reducing faces to line drawings abolishes holistic face processing. Holistic face processing during the composite face effect is evidenced by two faces fusing into a Gestalt when the top half of one face is horizontally aligned with the bottom half of another. Zhao and colleagues therefore argued that the 3D information in the texture and shading is essential for holistic processing. Whereas the 3D cues might be essential for holistic processing, the configuration of familiar faces can be distorted somewhat through stretching without any effect on recognition (Hole et al., 2002, see also Burton et al., 2015).

The composite face effect is reduced, however, by rigid rotations in yaw during the learning phase in which the motion retains the temporal sequence and fluidity (Xiao et al., 2012), suggesting that rigid motion might induce more part-based processing. This is consistent with evidence that holistic processing of unfamiliar faces is disrupted by changes in view (Carbon & Leder, 2006). It is also consistent with featural processing by AM and AF neurons in macaques (Waidmann et al., 2022), but conflicts with evidence that configural information is used more across view than featural information (Schwaninger et al., 2007), although both can be used reasonably effectively. But, the conditions were vastly different; the ‘configuration only’ condition consisted of blurred greyscale images and the ‘feature only’ condition consisting of a spatially segregated and scrambled face, so it is hard to directly compare the contribution of each.

Given that faces are nearly always in motion, the dependence on features (Xiao et al., 2012) could argue against the use of a single 3D internal model, although recent replications have failed to find this effect of motion unless presented over a large rotation from -90° to +90° (Y. Zhou et al., 2021). That said, Zhou and colleagues (2021) also found a reduced composite face effect when the viewpoint differed across the learning and test images compared to when presented from the same view or array of views as during learning. This discrepancy suggests holistic processing involves view-specific template matching rather than a 3D model. There was still evidence of holistic processing across views, however the authors described this as an activation and comparison of view-specific representations rather than using a 3D representation (Y. Zhou et al., 2021), consistent with the view-interpolation account.

It is plausible that representations are computed as a set of view-specific 3D representations rather than being integrated into one 3D Gestalt. One might refer to this as 2.5D processing, wherein the 3D representation is restricted to the specific viewpoint (Marr, 1982, as cited by H. Hill & Bruce, 1993), much like pin art toys where pushing your hand onto a board of pins recreates the shape on the other side. Such a representation has previously been used in DCNNs that ultimately achieve view-invariance (e.g., Yildirim et al., 2020) and other models have been created solely to recognise faces with 2.5D representations (Chong et al., 2019). Two and half dimensional processing has also been used to explain the inversion from a concave face mask to a convex face percept in the hollow face illusion (H. Hill & Bruce, 1993).

Troje and Kersten (1999) made a similar suggestion after finding that humans recognise profile images of themselves much slower than frontal. They suggested that object-centred representations might operate within the limited range of the input. In other words, an identity-specific 2.5D representation spanning all seen views, rather than either a full 3D or view-based model. If an identity-general 3D representation was present, one would expect that between the wide exposure to other faces, including siblings, and high familiarity with frontal and ¾ views of one's own face, we would have sufficient experience to efficiently apply and rotate a 3D representation for our own face. Instead, this 2.5D model blurs the lines between viewer-centred and object-centred representations, yet storing multiple, view-specific 2.5D representations is likely more computationally expensive than a single 2.5D or 3D model. It is possible of course that the brain uses both 2/2.5D template matching and 3D rotation systems, and that the increase in reaction times to one's own profile (Troje & Kersten, 1999) could be attributed to moving from faster template matching procedures to slower 3D representations. However, this implies familiar faces are predominantly processed using view-based matching, raising the question of why a 3D representation would be necessary.

There is also little behavioural evidence whether stereoscopic viewing, which enhances the perception of 3D structure (e.g., see McIntire et al., 2012), aids view-invariant recognition. As binocular vision is necessary for stereopsis, individuals who are monocularly blind from an early age have the potential to provide useful insight. Monocularly blind patients are slower at responding to featural and configural changes than controls and show impaired holistic processing (Kelly et al., 2012) suggesting that view-invariance may also be affected, but we have not yet seen any direct evidence for this. The use of symmetry is thought to be a stepping stone to view-invariance (e.g., Flack et al., 2019; Meyers et al., 2015; Rogers & Andrews, 2022) and while patients who are congenitally blind in one eye are slower at detecting symmetry in patterns, they match controls in accuracy (Cattaneo et al., 2014).

The depth perception lost through stereopsis in these individuals is thought to be compensated through the use of optic flow through voluntary head movements (Gao, Huang, et al., 2023; Gao, Liu, et al., 2023), which might allow the generation of a 3D representation. On the other hand, 2D processing does not necessarily require depth perception when learning view-based representations, so even without any compensatory strategies it should be relatively unaffected by monocular blindness.

In addition to some behavioural deficits, patients with early monocular enucleation also show reduced responses to faces in bilateral OFA and left FFA, but not in the right FFA (Kelly et al., 2019). It would be beneficial to also assess whether responses are modulated in the more view-invariant, anterior regions or whether responses to dynamic stimuli differ in, say, the pSTS. Evidently, face processing is somewhat disrupted in monocularly blind individuals, but it remains unknown if and how view-invariance is affected.

In binocularly sighted individuals, stereoscopic stimuli can be presented artificially by presenting images of slightly different viewpoints to each eye, mimicking stereopsis in natural viewing and creating apparent 3D structure. While recognition accuracy is improved during stereoscopic viewing (Chelnokova & Laeng, 2011; Eng et al., 2017), Eng and colleagues (2017) found marginally but not significantly reduced RTs whereas Chelnokova and Laeng (2011) found significantly increased RTs suggesting that while possible, 3D representations might come at a computational cost. Accuracy for inverted stimuli, however, was not increased for 3D over 2D stimuli, suggesting that the benefit for upright faces viewed stereoscopically could not simply be explained by the presence of additional depth information and instead that 3D representations tap into face-selective processes. Despite this, three quarter-frontal matching was marginally better than frontal-frontal but did not interact with 2D vs 3D viewing (Chelnokova & Laeng, 2011) so it is unclear whether 3D viewing actually aids recognition across views. Likewise, for familiar individuals there is conflicting evidence whether processes rely more on 3D structure (see Blauch & Behrmann, 2019) or surface properties that could be extracted from 2D representations (Russell & Sinha, 2007). Afterall, the neural responses in anterior face patches in macaques are more sensitive to texture and surface properties than shape (Chang & Tsao, 2017). Recent technological advancements, particularly in VR, should allow more elaboration on the 2D/2.5D/3D representation discussion, although the results thus far are not conclusive (see Burt & Crewther, 2020).

Stereoscopic viewing during fMRI has been found to increase the BOLD response in the OFA, but not the FFA (Deligiannis et al., 2023) suggesting that the OFA might represent faces in a 3D manner. However, the non-significant effect in the FFA leaves outstanding questions. Does the FFA process faces in 3D? If so, why was there no significant effect of stereoscopic viewing? Is it possible that the FFA is so efficient at estimating 3D structure that it can do so from 2D displays as easily as stereoscopic displays? Or does the FFA collapse information into 2D view-dependent representations? As noted by the authors, more work should be done on stereoscopic viewing during fMRI to elaborate on their findings.

An illusion that is stronger when not using stereopsis is the hollow face illusion. When first seeing the hollow face illusion (Gregory, 1973, as cited by H. Hill & Bruce, 1993) one likely assumes that a 3D representation must be present. In this illusion, the perceived depth of a concave mask inverts eliciting a convex percept that follows you around the room. From observation, it appears that the illusion must indicate a single 3D representation. Yet, studies investigating influences on this illusion suggest this likely is not the case.

Firstly, the illusion appears at a shorter distance when viewing monocularly compared to binocularly (H. Hill & Bruce, 1993), showing that the illusion is driven by cues available within a 2D view, such as shape-from-shading and an assumption of convexity. It can only be overridden by binocular depth cues to a limited degree, however, as the illusion still holds at further distances. The assumption of convexity seems to be more selective for naturally observed faces; upright masks are falsely perceived as being convex more than inverted masks (H. Hill & Bruce, 1993), as are naturally textured and coloured stimuli compared to grey stimuli (H. Hill & Johnston, 2007). The emergence of the illusion is also dependent on the lighting condition (H. Hill & Johnston, 2007) indicating that both upright faces and appropriate lighting angles are important. Overall, the dependence on orientation, texture and lighting speaks against a full 3D representation of faces, in favour perhaps of 2.5D representations that are more dependent on viewing conditions (H. Hill & Bruce, 1993).

Although not sufficient for object recognition, evidence from individuals with object recognition impairments suggests the dorsal visual route holds a coarse description of 3D object shape (Freud et al., 2017). Both controls and patients exhibited increased fMRI responses in dorsal regions to shapes with impossible 3D structure. A coarse description of 3D facial structure might also exist. Although, it raises the question of whether the 3D representation is used for view estimation or view-invariance, or whether it simply signals unlikely stimuli given prior experience. Given the cross-talk between ventral and dorsal regions in object processing (Freud et al., 2017), it is possible that dorsal regions might communicate with ventral face-selective regions for faces that are more implausible.

Perhaps some unexpected evidence for a 3D representation comes from results showing that the FFA also responds to the haptic exploration of 3D-printed faces in the congenitally blind (Murty et al., 2020) and in seeing participants trained on haptic exploration (Kilgour et al., 2005). During haptic exploration it would be possible to discriminate viewpoint but having view-dependent receptors does not seem appropriate for this form of sensory input, instead a mental 3D representation seems more plausible. Kilgour and colleagues (2005) only found increased activation in the left FFA for faces over control objects during haptic exploration and suggested this was due to the necessity to process features sequentially, as the left FFA has been argued to process face parts (Rossion et al., 2000). However, these participants were only trained on haptic exploration for a relatively short time. Congenitally blind patients in contrast, who have more experience, showed face-selective responses in the right FFA (Murty et al., 2020).

As the right FFA is thought to process faces holistically (Harris & Aguirre, 2010; Rossion et al., 2000) the results may indicate the construction of a whole-head 3D representation. Of course, between the lack of visual surface cues and the ability to use proprioception, haptic exploration might motivate and allow for more 3D structural representations in, what may inherently be, a 2D visual system.

Interestingly, the temporal dynamics of object view-invariance in macaque IT depends on the object's 3D structure and how its projections onto the retina change with view. Neurons are faster to respond in a view-invariant manner to objects that do not change much in aspect ratio compared to those that do (Murty & Arun, 2015). This poses a challenge for a 3D representation of objects unless all objects are processed with respect to one spherical template, which seems maladaptive given the variety of object structures in nature. Faces overall do not change much in aspect ratio as a function of yaw, but, as noted by Favelle and colleagues (2017), the aspect ratios of individual features can change substantially. Some such features, such as the nose, also provide volumetric information and receive greater levels of attention under stereoscopic viewing than 2D viewing (Chelnokova & Laeng, 2011), although other volumetric cues that do not change much in aspect ratio such also the cheek also receive more attention.

Research showing a behavioural advantage for specific views might also provide insight as to whether 2D or 3D representations are used. Despite faces being viewed from almost every angle, there is behavioural evidence for a ¾ view advantage (Marotta et al., 2002; O'Toole et al., 1998; Troje & Bülthoff, 1996; Van der Linde & Watson, 2010) suggesting this might be one of the canonical views that the brain represents if processing faces in a 2D view-based manner. In contrast some studies have shown no advantage for ¾ over frontal views (Carbon & Leder, 2006; Favelle & Palmisano, 2018) and others a frontal view advantage (Favelle et al., 2017; H. Hill et al., 1997) suggesting that frontal views might too be one of these preferentially represented views, consistent with SUR studies in macaques (Perrett et al., 1991). Interestingly, Van der Linde and Watson (2010) showed the ¾ view advantage at both learning and test when participants had to recognise identities across changes in view, but a frontal view advantage when faces were seen at the same view in learning and in test. Therefore, the ¾ view might be especially informative for achieving view-invariance. The ¾ bias could be due to the visibility of transformation points for performing mirror-symmetric computations, that ¾ views provide the most information about 3D shape or that they might contain the most diagnostic information about identity (H. Hill et al., 1997; Marotta et al., 2002; Troje & Bülthoff, 1996). Ramírez et al (2014) found a larger FFA response to frontal faces over both ¾ and profile faces, although this could reflect template matching procedures whereas more anterior regions might be more sensitive to ¾ views for achieving view-invariance.

While frontal and ¾ views have been compared relative to each other, few studies have systematically compared them to intermediate views (e.g., 22.5^o) to shed light on whether these views are preferentially represented. Chapter 6 therefore describes a study we conducted assessing whether some views are preferentially represented over others.

Overall, it remains unclear whether faces are processed via one 3D model or via multiple view-specific 2/2.5D representations. From the evidence presented above it seems more likely that a view-based interpolation account of face processing is used, rather than a 3D representation. This is particularly suggested by the evidence that the brain can be tricked into thinking multiple different identities are one when presented in quick succession and when presented as a smooth transition across views (Wallis, 2002; Wallis & Bülthoff, 2001). A coarse description of 3D shape is likely present, and depth cues can be used, however the results overall suggest a 2D or 2.5D view-based system rather than a 3D system.

Rather than investigating behavioural and cortical responses to find evidence for or against 3D representations, we focused on whether it was feasible to create a 2D view-based model in a biologically plausible manner. This PCA-based, multi-view face space model, which builds on the work by Beridze (2021), might give some insight into whether a 2D view-based system is possible in humans.

Chapter 2 An overview of PCA

2.1 Preface

As outlined in the previous chapter, principal component analysis (PCA) has often been used for constructing computational models of face space, among many other computer applications. In this chapter, we provide more detail on the underlying mathematics, how it can be used to create face spaces, and how images can be recovered from it.

2.2 Performing PCA

PCA was first developed over 100 years ago by Karl Pearson (1901) as a method for fitting lines and planes to data. The technique is a data-driven method taking high dimensional data and returning a new set of orthogonal basis vectors that are a linear combination of the initial dimensions. The new dimensions try to fit the data as closely as possible and are ordered based on the amount of variance they explain. Dimensions coding only a small proportion of the variance can be ignored, hence PCA is often used for dimensionality reduction. Because PCA is guided by the covariance in the data and not on external predictions it is excellent for exploratory analyses.

One of the assumptions of PCA is that the data is zero/mean-centred. For a matrix X, the mean vector $\bar{X}$ is first calculated and removed. $\bar{X}$ will have as many features as is in the data. The orientation of X seems to vary depending on the usage, but for the implementation of PCA discussed below (singular value decomposition) it has n

columns which reflect different samples and m rows which reflect different features, e.g., pixels in a vectorised image.

The process of PCA can be thought of iteratively. The first dimension should explain the highest possible amount of covariation in the data. The orientation of this line is calculated in a similar way to linear regression by minimising the sum of squared distances to the line, but not directly. The calculation can also be performed with projection. The projection of a data point (a) onto a line (b) is the point at which a line drawn from a intersects b at 90°. The aim is to find the line of best fit that maximises the sum of squared distances $(\sum d^2)$ between the intersection points and the origin (d in Figure 2.1a), and in doing so minimises the distances between the data points and the line. Note that this line can be at any orientation but must pass through the origin. Figure 2.1a shows an example of this, where vector i_1 is iteratively rotated n times until $\sum d^2$ is maximised. The next component can be calculated by performing the same series of calculations, but with the constraint that the vector must be orthogonal to the prior component(s). One method is to subtract out the previous dimensions and calculate the next that best fits the residual data.

The eigenvectors or singular vectors are simply the new dimensions, normalised to have unit length (a length of 1, Figure 2.1b). Each eigenvector is a linear combination of the original feature dimensions, with each element in the vector expressing the relative contribution of each original feature. I.e., principal component (PC) 1 $=$

$\beta_1 a + \beta_2 b + \beta_3 c$ where a, b and c are the original dimensions or features and β_1, β_2 and β_3 describe the weights. In this sense it can be thought of as a rotation relative to the original dimensions.

The eigenvalue and singular value for an eigenvector can be calculated as $(\sum d^2)/(n-1)$ and $\sqrt{}(\sum d^2)$ respectively. The eigenvalues are a measure of variation, and thus the percentage of variance explained by each component can be calculated. Say we have three components, PCs 1, 2 and 3, and the eigenvalues are 31, 12 and 8. The total variance is 51. The variance explained by PC 1 is 60.8% (31/51 x 100).

Once the explained variance of later (often called higher) components drops below a threshold, be it arbitrary or based on the other components, the subsequent components can be ignored. For example, in our multi-view PCA models, n components are retained such that $n \geq 100$ and a minimum of 90% of the variance is explained. Other methods for example plot the variance explained in a scree plot, which often reveal a sharp initial decrease in the variance explained, followed by a plateau. There may be a deflection point at which this plateau occurs, which is used as the cutoff.

Calculating PCA by hand is obviously going to be time-consuming and is not feasible for most datasets, and the same is true even for a computer when doing an iterative search. Instead, mathematical and computational solutions exist. One such way is through singular value decomposition (SVD). While the process for how SVD is calculated is beyond the scope of this thesis, some of the main points of interest are highlighted. There are many webpages, books and YouTube videos that can explain PCA and SVD in more detail, such as "Data-driven Science and Engineering: Machine Learning, Dynamic Systems and Control" (Brunton & Kutz, 2022) or the associated YouTube videos by Professor Steven Brunton.

SVD is generally written in the form $X = U\Sigma V^T$, and the MATLAB implementation is simply `` `[U,S,V] = svd(X);` ``. X is the original data matrix. U is the matrix of eigenvectors that we might refer to as eigenfaces, eigenframes (EFs) or principal components (PCs), in our context. It has a column for each component and a row for each element or feature. The number of components is generally truncated, either to the specified number or to the number of the columns in the input matrix. This is because it cannot have more linearly independent columns than there are samples. Σ is the diagonal matrix of singular values, and this too is generally truncated. V is another matrix of eigenvectors, but these explain how to combine the eigenvectors in U to recreate the vectors in X, once scaled by Σ. Each row in V explains how to combine the columns of U to make one frame or image.

In relation to Figure 2.1a and b, U describes the rotation of the components in relation to the original features. For example, the original features in Figure 2.1 might be the values for the first 2 pixels in an image. U therefore describes the weighted combinations of pixels 1 and

2 needed to make unit vectors in the direction of each column of $U.$ In Figure 2.1b the solid, shorter red arrows depict the first two columns of U. Note that the values within a component reflect one polarity, e.g., the arrow pointing up and right from the origin in Figure 2.1b. To invert the polarity and therefore the direction of the arrow relative to the origin, the components are multiplied by -1.

Σ then defines a stretch. Each column in U is scaled by its diagonal element in $\Sigma.$ Because Σ is a diagonal matrix, this can simply be computed with matrix multiplication. Finally, V^T describes another rotation detailing how to transform from the scaled components in U back to the original data. The principle is the same as for U. The matrix V has a column for each component in U and a row for each image. It describes what ratio of the scaled components are needed to reconstruct each image.

Essentially, U describes the rotation to make the new basis vectors, Σ indicates how much of the variance each component explains and stretches U, and V describes how to rotate the new stretched dimensions back to the original data.

One additional consideration to note is how SVD relates to covariance matrices of X. Although rarely used due to the complexity of the task when working with lots of elements, the principle and the outcome is essentially the same.

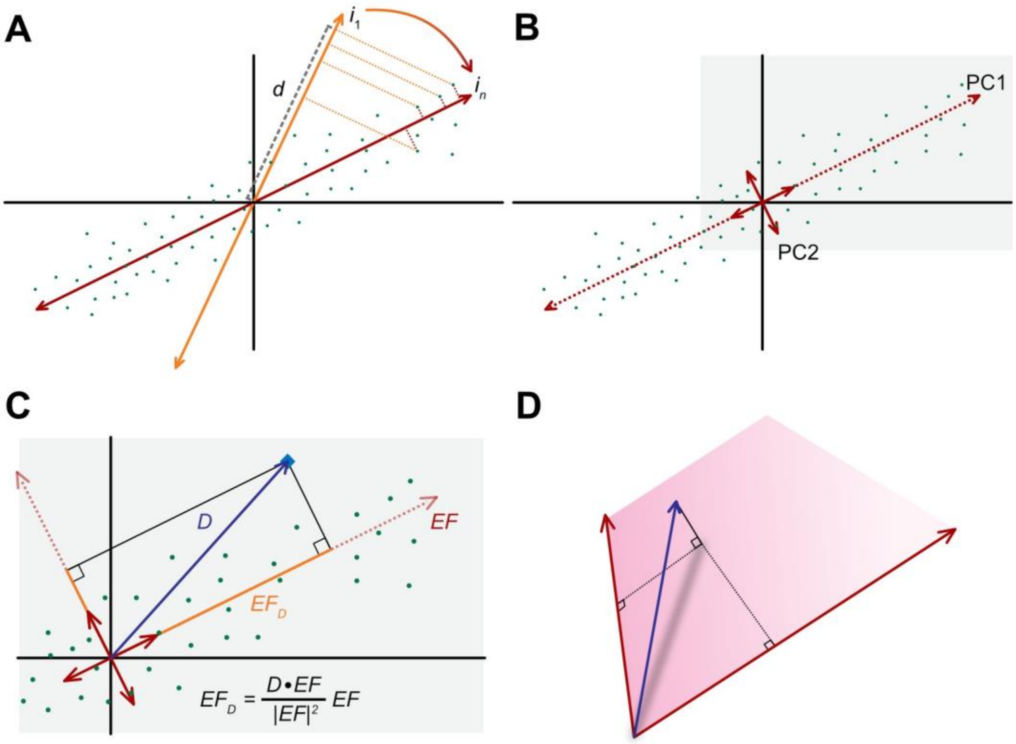


Figure 2.1. PCA space and projection.
(**A**) Calculation of PC1. A cloud of data points (green dots) against the original dimensions (black horizontal and vertical lines). Rotation of an initial estimate for PC1 (orange, i_1) to its final position (red, i_n), such that the $\sum d^2$across all data points is maximised. The orientation of these lines describes the eigenvector, the length of vector is proportional to the eigen and singular values. (**B**) The first and second components (*PCs*), also known as eigenvectors or in our case eigenframes (EF). Solid arrows are scaled to be unit vectors. The grey box outlines the field of view for **C**. (**C**) Projecting the new data point, in our case a deviation vector, D (blue diamond/line) onto the components, alongside equations showing how the vector EF_D is calculated. (**D**) Example of a vector in 3D space projected onto a 2D hyperplane and its dimensions.

Firstly, we know the equation for SVD, but how do we find out the values for U, Σ and V? To find the solution for V we first need to cancel out U, and vice versa to find the solution for U. As it happens, U and V are orthogonal matrices of unit length, so $U^T U$ and $V^T V$ are both equal to an identity matrix (a square matrix of zeros with ones down the diagonal), which cancel out within the following equations. Now, the transpose of X (referred to as X^T) equals $V\Sigma^T U^T$, so U or V can be cancelled out by multiplying X by X^T depending on which way around the multiplication occurs. $X^T X$ is the covariance matrix of the columns of X. Each element is the inner product between the two columns of X, X_i and X_j where i and j refer to the row and column in the covariance matrix respectively. Expanding we have, $X^T X = V\Sigma^T U^T U\Sigma V^T$. As $U^T U$ is an identity matrix this term cancels out and because Σ is a diagonal matrix it multiplied by itself is Σ^2, $\therefore X^T X = V\Sigma^2 V^T$. The solution for V is the eigendecomposition of the covariance matrix for the columns of X.

To find U, both sides of the equation are instead right-multiplied by X^T which is now the covariance matrix of the rows of X. $XX^T = U\Sigma V^T V\Sigma U^T$. Again, $V^T V$ cancels out so the equation can be reduced to $XX^T = U\Sigma^T U^T$. U can therefore be determined through the eigendecomposition of the covariance matrix for the rows of the X (XX^T). The eigenvalues Σ will be the same for both the eigendecomposition of $X^T X$ and XX^T. The reason for not performing this calculation directly is because the covariance matrix (XX^T) will have as many rows and columns as X has rows. In some of our uses below, X has over 400,000 rows. MATLAB would need over 1TB of memory just to hold the corresponding covariance matrix of type ‘double’.

While we won’t go into the details of how the calculation works, the important part is that the calculation is an approximation for the above. This is important because the covariance matrix multiplies everything by itself down the diagonal, so all numbers down the diagonal are positive, irrespective of the sign of the polarity of the data. As a result, the ‘negative’ and ‘positive’ directions of the components are arbitrarily labelled, they are simply opposing deviations away from the mean.

2.3 Projecting into the PCA space

After performing PCA on dataset X, the next task is to assess how new data relates to our newly defined components. This is done through projection. A depiction of both PCA calculation and vector projection can be seen in Figure 2.1. Mathematically, projection finds the point on a vector (EF) that minimises the distance between vector D and vector EF. Visually, this forms a right-angle triangle (Figure 2.1c).

To project a vector V into the space, the mean vector ($\bar{X}$) from the PCA construction is first subtracted to give a deviation vector (D). D is then projected onto the eigenvectors using inner products (also known as dot products or scalar products). This provides one projection loading per input vector and eigenvector.

The full equation for the projection of a vector a onto a vector b ($Proj_b a$) can be summarised by equation (2.1). In this equation, the output is a vector in the direction of b which is $\frac{a \cdot b}{|b|^2} x$ the length of b, where $\frac{a \cdot b}{|b|^2}$ is a scalar. The final version of the equation shows how it would be written as a matrix operation, where ' is the transpose.

$$Proj_b a = \frac{a \cdot b}{|b|^2} b = \frac{a \cdot b}{b \cdot b} b = \frac{a'b}{b'b} b \qquad \textbf{(2.1)}$$

In the context of our deviation vectors and eigenframes, a = D and b = EF. $Proj_{EF}D$ returns the vector EF_D in Figure 2.1c. EF_D is in the direction of EF but is $\frac{D \cdot EF}{|EF|^2} x$ the magnitude. In this case, $\frac{D \cdot EF}{|EF|^2}$ is the loading of D onto $EF.$

The magnitude of this loading is calculated by the inner product of D and EF, divided by the squared length of EF, which is the inner product between EF and itself. The inner product is simply the sum of the element-wise products. See equation (2.2) where e is the number of elements in the vectors.

$$D \cdot EF = \sum_{i=1}^{e}(D_i EF_i) \qquad |EF|^2 = \sum_{i=1}^{e}(EF_i)^2 \qquad \textbf{(2.2)}$$

Often, PCA uses singular value decomposition (SVD), which returns normalised eigenvectors. These vectors have a length of 1 (i.e., are unit vectors) and so the denominator in the projection equation is 1, and thus can be ignored (when $|EF| = 1,\ D \cdot EF = \frac{D \cdot EF}{|EF|^2}$).

The equation for projecting multiple frames onto a given component EF_i is provided in equation (2.3), where i is the component number, n is the number of frames, and e is the number of elements in the vector. The resulting output, L, is an $n\ x\ 1$ column vector of loadings. When projecting onto multiple components, the outputs can be concatenated to provide a single $n\ x\ p$ matrix where p is the number of components. Again, if the vectors in EF are unit vectors then the denominator can be omitted.

$$L \ = \ \frac{F_i}{EF_i \cdot EF_i} \quad where\ F_i = \begin{bmatrix} D_{1,1} & \cdots & D_{1,n} \\ \vdots & \ddots & \vdots \\ D_{\mathrm{e},1} & \cdots & D_{e,n} \end{bmatrix}' \begin{bmatrix} EF_{1,i} \\ \vdots \\ EF_{e,i} \end{bmatrix} \qquad \textbf{(2.3)}$$

In MATLAB the above can be calculated for all components simultaneously using element-wise division in the following:

`(D'*EF)./diag(EF'*EF)`

At this stage we have constructed a PCA space and calculated the loadings of a new set of data, or videos frames on the components. When scaled by Σ, the matrix V in SVD is a matrix of loadings. The next stage, often of interest, is reconstructing an estimate of the input.

2.4 Reconstructing from the PCA space

A good way to assess the quality of the information stored in a PCA space is to reconstruct a vector, image or video projected into the space. If the space satisfactorily encodes data contained in the input vector then it should be able to reconstruct it well.

A given frame or vector (D) can be reconstructed by first projecting it into the space (EF) as in the previous section. The loadings are then used to scale each component and for each input vector the scaled components are then summed. This gives a reconstructed deviation vector (D_r). Finally, the mean vector $\bar{X}$ is added. Each reconstruction is essentially a weighted sum of the components + $\bar{X}$. The resulting vector can then be reshaped or warped back to an image.

Equation (2.4) describes how to reconstruct D_r from D, where e is the number of elements in the vectors and components, and p is the number of components. D_r is a column vector.

$$D_r = \begin{bmatrix} EF_{1,1} & \cdots & EF_{1,p} \\ \vdots & \ddots & \vdots \\ EF_{e,1} & \cdots & EF_{e,p} \end{bmatrix} \begin{bmatrix} L_1 \\ \vdots \\ L_p \end{bmatrix} = \sum_{i=1}^{p} (L_i * EF_i) \qquad \textbf{(2.4)}$$

To reconstruct multiple images, a matrix of loadings is used instead of a vector of loadings, as equation (2.5). In L there is a row for each component and a column for each frame. Note that this is transposed relative to the output of the projection process. D_r is an $e\ x\ n$ matrix. Again, $\bar{X}$ is added to the reconstructed vectors.

$$D_r = \begin{bmatrix} EF_{1,1} & \cdots & EF_{1,p} \\ \vdots & \ddots & \vdots \\ EF_{e,1} & \cdots & EF_{e,p} \end{bmatrix} \begin{bmatrix} L_{1,1} & \cdots & L_{1,n} \\ \vdots & \ddots & \vdots \\ L_{p,1} & \cdots & L_{p,n} \end{bmatrix} \qquad \textbf{(2.5)}$$

This reconstruction procedure can be used with loadings calculated using full vectors, or it could be performed with partial vectors. In this case, the portion of interest is retained and the unwanted portions of the input vector are replaced with zeros. Because the full components contain information other than in the portion of interest, it acts as an auto-associative memory and allows, for example, the whole face to be recovered from part of the face (Berisha et al., 2010; Turk & Pentland, 1991) or indeed for different modalities to be recovered from each other (Scholes et al., 2020).

2.5 Face space representations

As outlined in Chapter 1, PCA is a popular choice for making computerised versions of face space, but there are many possible options for making such a space. Here we will outline just a couple of methods that are most directly relevant for the multi-view PCA work described later.

Early PCA-based spaces simply used vectorised images as inputs to the space (Sirovich & Kirby, 1987; Turk & Pentland, 1991). The eigenvectors returned describe deviations in the pixel intensities (see Fig 2 on page 75 of Turk and Pentland). An image could be reconstructed by projecting it onto the eigenvectors and calculating a weighted sum of the eigenvectors based on the loading of the projection. While this method worked, it resulted in substantial issues with superimposition. By using raw images, the PCA was agnostic to any spatial differences caused by movement or differences in face shape. As a result, changes say in the width of the face were simply encoded by superimposing wider cheeks over the images rather than directly changing the shape of the face.

More sophisticated methods separate out shape and texture information which allow for changes in shape. These active appearance models (AAM, Cootes et al., 2001; Edwards et al., 1998) first align the shapes of the faces to give a ‘shape free’ (or rather shape normalised) texture map (or appearance). For instance, Burton and colleagues (2016) placed 82 landmarks onto various images of near-frontal faces, which were then used to warp the shape of each input image to and an average template. The locations of the landmarks prior to warping are also stored. This provides two descriptors for each image: one describes the texture (RGB values) once spatially aligned to the template, and the other describes the spatial properties of the face prior to warping. In this case the spatial properties are encoded by two vectors: one containing the horizontal locations of the landmarks, and one the vertical locations. This pipeline has since been incorporated into the freely available InterFace software (Kramer et al., 2017), written in MATLAB. Both of the shape information and the texture information can then be reshaped into vectors. Often, these shape and texture vectors are entered into separate PCA analyses (Andrews et al., 2023; Burton et al., 2016; Chang & Tsao, 2017). Reconstructions are then made by summing the weighted components in each space and then warping the summed texture vector by the shape in the summed shape vector. Even just by normalising face shape, Burton and colleagues (1999) found a 12% increase in their model's recognition accuracy compared to simply aligning images by eye position.

While separating shape and texture into separate spaces can be helpful, it is not always necessary. One reason for separating them is that the vectors have different units and therefore the magnitudes differ. If combined, PCA can be biased in its estimates of the major sources of variation to the modality that has the greater variance in magnitude. Despite this, many studies have concatenated shape and texture deviations (Beridze, 2021; Cowe, 2003; Nagle et al., 2013; Scholes et al., 2020; D. M. Watson & Johnston, 2022). Whilst a valid concern, it is also not one that appears to have dramatically affected our multi-view PCA work. We would likely see slightly different components if shape/motion and texture were separated, but the reconstructions seem accurate and seem to vary in both. The necessity to separate them may depend on the variability in the images. In our work the videos were captured in controlled conditions and factors such as lighting did not vary, so there was no need to separate out shape changes from say the effect of illumination. Separating texture and shape would have also led to additional complexity given that we were trying to combine motion across multiple viewpoints. Furthermore, combining shape and texture allows covariations between texture and shape to be extracted.

To establish the shape features, many options are available. As described above, one could use a relatively small number of landmarks either placed manually or automatically using software such as OpenFace (Baltrusaitis et al., 2018). Alternatively, one could use warping methods that are agnostic to facial landmarks, such as the Multi-channel Gradient Model (McGM, Johnston et al., 1992, 1999). The McGM provides a much denser description of changes in shape, output as pair of warp fields each with a value for every pixel in the image. One outlines horizontal spatial information, the other vertical. More information about the McGM is provided in Chapter 4.

Once the shapes have been aligned through landmarking, warping or otherwise, the interesting applications of PCA can begin, such as passing information between spaces and learning covariations across different datasets. Below we have expanded on some of the uses mentioned in Chapter 1.

Firstly, PCA can be used to learn the covariations across different datasets and modalities. For instance, Scholes et al (2020) learned how the external appearance of the face covaried with the internal structure of the vocal tract. Participants read out sentences in a recording studio, and separately during an MRI scan. Scholes et al then concatenated vectorised versions of the video frames with vectorised MR slices of the vocal tract. Both contained X/Y warp fields and texture deviations of the frames relative to the modality's average. By concatenating the vectors that described these different modalities and entering the combined matrix into a single PCA they were able to assess how external changes in the face correlated with internal changes in the vocal tract. These covariations were captured in the components and thus the auto-associative memory of PCA allowed Scholes and colleagues to reconstruct the appearance of the vocal tract from the external image of the face and vice versa. The approach is equivalent to that used by Beridze (2021) to reconstruct facial motion across different views and is also the approach we use to expand Beridze's work in Chapter 4.

Information can also be mapped between separate spaces (Cowe, 2003; Griffin et al., 2011; Nagle et al., 2013). Griffin and colleagues constructed two PCA spaces, one for female and one for male faces. They then took an image of a male face, say, and calculated the deviation from the average male image. The deviation vector was then projected into both spaces, simply calculating the inner product between the vector and the components. The weighted sums of the components were calculated and the average image for each gender added. These reconstructed images shared familial resemblance as the deviation vector loaded onto components expressing a similar change in the female space. These need not be the numerically corresponding components. If PCs 1, 3 and 6 were heavily loaded onto in the male space it would *not* have to be these components that were loaded onto in the female space.

Similar methods were used by Cowe (2003) and Nagle and colleagues (2013) to map facial expressions and motion across different identities. Identity-specific spaces were first made by performing PCA on multiple vectorised images for each identity, with each input vector again containing X/Y warp fields and RGB texture information. In this case the origin was the mean facial representation for the given identity, and the components and deviation vectors coded changes in expression. A deviation vector for given frame for one person could be projected into their space and into the spaces of others, loading heavily onto components containing similar information and thus reconstructing similar expressions.

In all cases (Cowe, 2003; Griffin et al., 2011; Nagle et al., 2013), however, there are limits to what can be reconstructed. Information can only be reconstructed if it was on the span of information included in the training set. In the example in Figure 2.1d, the PCA space only spans two dimensions, and so any information the test vector might have about a third dimension is lost. In the case of face spaces, males vary in the presence and appearance of facial hair and so this will be coded in at least one component. The female space is much less likely to have any such experience, and so no components would represent facial hair. The reconstruction of a bearded individual from the female space will no longer have a beard. Likewise, expressions can only be reconstructed if similar expressions have been seen in both identities, or if they can be created through a combination of other expressions.

As will be discussed more in Chapter 4, the fidelity of reconstruction also depends on the spatial correspondence when mapping between spaces. The cross-space reconstructions described (Cowe, 2003; Griffin et al., 2011; Nagle et al., 2013) work because the faces are all frontal and have sufficient spatial overlap through an affine transformation to align the positions of the eyes and the mouth.

As well as reconstructing best approximations of input images one can also manipulate the weights of the loadings to change the appearance of a face. Once the loadings of an input face on the components are known, they can be increased or decreased to caricature or anti-caricature respectively. They can, for example, also be multiplied by -1 to make an anti-face (e.g., Blanz et al., 2000; Leopold et al., 2001), or -0.5 to make an anti-caricatured anti-face.

Similarly, appearance can be manipulated by changing the direction of an image in face space through changing the relative loadings on the components, or indeed certain components can be removed entirely. For example, by selectively ignoring components when making the reconstructions, Andrews and colleagues (2023) found a narrow band of components crucial for recognition.

As discussed in the previous chapter, the orthogonality of PCA is particularly interesting in light of recent evidence in macaques showing that neural responses in the face patch AM are unaltered by changes orthogonal to the information the neuron preferentially responds to (Chang & Tsao, 2017). This ambivalence to orthogonal axes allows face processing to occur as a linear projection as in PCA.

The next chapter will explore what happens and how face- and object-selective areas of the brain respond when loadings on the components in a face space are scaled beyond the realm of natural plausibility.

Chapter 3 fMRI evidence that hyper-caricatured faces activate object-selective cortex

3.1 Preface

This chapter contains the published manuscript for an fMRI study we conducted looking at the BOLD response to faces over varying caricature level. Other than matching the format to the rest of the thesis, the manuscript has not been altered. There is therefore substantial overlap between in the introduction for this chapter and the main introduction. The paper has been published in Frontiers in Psychology and can be found here: https://doi.org/10.3389/fpsyg.2022.1035524.

3.2 Abstract

Many brain imaging studies have looked at the cortical responses to object categories and faces. A popular way to manipulate face stimuli is by using a "face space”, a high dimensional representation of individual face images, with the average face located at the origin. However, how the brain responds to faces that deviate substantially from average has not been much explored. Increasing the distance from the average (leading to increased caricaturing) could increase neural responses in face-selective regions, an idea supported by results from non-human primates. Here, we used a face space based on principal component analysis (PCA) to generate faces ranging from average to heavily caricatured. Using functional magnetic resonance imaging (fMRI), we first independently defined face-, object- and scene-selective areas with a localiser scan and then measured

responses to parametrically caricatured faces. We also included conditions in which the images of faces were inverted. Interestingly in the right fusiform face area (FFA), we found that the patterns of fMRI response were more consistent as caricaturing increased. However, we found no consistent effect of either caricature level or facial inversion on the average fMRI response in the FFA or face-selective regions more broadly. In contrast, object-selective regions showed an increase in both the consistency of response pattern and the average fMRI response with increasing caricature level. This shows that caricatured faces recruit processing from regions typically defined as object-selective, possibly through enhancing low-level properties that are characteristic of objects.

3.3 Introduction

In regular social interactions, we may encounter hundreds of faces every day. Most human observers can rapidly recognise the identity (Ramon et al., 2011), process the emotion (Leppänen & Hietanen, 2004), or form an impression of a person or their intentions (Bar et al., 2006; C. A. M. Sutherland et al., 2013; Willis & Todorov, 2006) from visual information alone.

It is estimated that humans know on average around 5000 faces (Jenkins et al., 2018) but despite much research, it is largely unknown how we encode all those familiar faces, in addition to all unfamiliar ones. Face space (Valentine, 1991; Valentine et al., 2016), an influential account of face representation, has been widely used to study the neural representation of faces in humans (Carlin & Kriegeskorte, 2017; Loffler et al., 2005) and non-human primates (Chang & Tsao, 2017; Leopold et al., 2006). The idea has also found application in automatic face recognition systems (Deng et al., 2014; Sirovich & Kirby, 1987; Turk & Pentland, 1991; Zhu et al., 2013). Caricatured face images that deviate substantially from average, including artistic caricatures, evidently amplify characteristic features of faces. But it is unclear if and how face space is represented in the brain and what the exact neural representation of faces distant from the average face might be.

In *face space*, a multidimensional space with a representation of an average face at the origin, individual face *exemplars* are thought of as points (at a certain distance and direction with respect to the origin). The dimensions of this space could be derived from discrete, descriptive changes in the shape or position of features (e.g., the distance between the eyes or the width of the mouth). Alternatively, the dimensions may reflect more abstract and global descriptors of shape and texture.

In a face space representation, individual *identities* correspond to a given direction relative to the origin. The distance from the origin indicates how different a particular face is from the average. Faces whose representation is located a greater distance from the average are expected to generate stronger responses from the population of neurons sensitive to the given identity’s facial properties. This idea of "norm-based” coding, coding relative to the average or norm, has received strong supporting evidence (e.g., see Anderson & Wilson, 2005; Chang & Tsao, 2017; Jiang et al., 2006, 2007; Leopold et al., 2001; Little et al., 2012; Rhodes & Jeffery, 2006; Webster & MacLeod, 2011). Neurons representing facial information could form a basis to span this space, rather than being tuned to a particular identity. The projections of a face onto a set of basis neurons may code the different identities (Chang & Tsao, 2017) in terms of the relative firing rates of this population of neurons.

The neural basis of norm-based coding has recently been clarified by new research in macaques (Koyano et al., 2021). Rhodes and Jeffery (2006) proposed that norm-based coding was based on two opponent channels with the average face activating each equally. The opponent channels can be associated with ‘axis coding’ that shows monotonic ramp-tuning through the norm in single-cell recordings (Chang & Tsao, 2017). Norm-based coding also gives rise to V-shaped coding (e.g., Freiwald & Hosoya, 2021) whereby there is a minimum response to the norm relative to more peripheral faces, regardless of direction. V-shaped coding was first demonstrated at a single cell level and at a population level by Leopold and colleagues (2006). Recently, evidence has shown that both mechanisms are present in the same set of individual neurons, with axis coding occurring approximately 100ms before V-shaped coding (Koyano et al., 2021). However, the V-shape was driven by a decrease in the firing rate to average faces, likely from lateral inhibition resulting from synchronous firing across the population to the average face (Koyano et al., 2021). While axis coding supports the initial coding of the neuron, V-shaped responses reflect a consequence of many neurons firing to the average face in synchrony.

Chang and Tsao (2017) show that rather than responding to specific identities, neurons in areas ML/MF of the macaque temporal lobe (middle lateral/middle fundus) and the more anterior AM (anterior medial) responded to combinations of shape and texture information. Firing rates increased linearly with the magnitude of a face's projection onto the neuron's preferred dimension or 'axis' of change, but only in the preferred direction of change; face stimuli along the same axis but on the opposite side of the mean decreased the neuron's firing rate. Variations in facial appearance orthogonal to a neuron's preferred dimension, however, did not change its firing rates. This invariance to changes along orthogonal axes may explain the lack of an aftereffect to faces that lie on a different trajectory from the adapting stimulus (Anderson & Wilson, 2005; Leopold et al., 2001; Rhodes & Jeffery, 2006). From a theoretical standpoint, it allows face processing to be based on a highly efficient calculation (linear projection), requiring relatively few neurons to encode a very high-dimensional face space. Interestingly, it has also been found that faces activate a more broadly-based representation within an object space. Recent work has shown that faces may be situated in the animate, 'stubby' quadrant of the identified 2D (animate/inanimate and stubby/spiky) space, although many aspects of this representation remain unknown (Bao et al., 2020).

Because faces are more densely clustered around the mean, those further from average should appear more distinctive (Valentine et al., 2016). This idea is supported by evidence showing that caricatures are rated as more distinctive than their veridical face or anti-caricature (K. Lee et al., 2000). If the dimensions are ordered in terms of the amount of facial variance they encode, then more distinctive faces may also load more onto less prevalent dimensions of variation, in which case direction in the space may also reflect distinctiveness (Hancock et al., 1996). The direction and distinctiveness in face space not only impacts recognition, but also the first impression that is attributed to that face (Olivola et al., 2014; Over & Cook, 2018), and can indicate poor childhood health or genetic disorders (Babovic-Vuksanovic et al., 2012; Dolci et al., 2021; Gad et al., 2008; Rhodes et al., 2001).

Faces can also be made artificially more distinctive through caricaturing. Caricatures, versions of “veridical” face images that can be derived from extrapolations in face space, enhance behavioural performance over veridical faces, suggesting they may elicit stronger responses in the brain. Caricaturing line drawings and photographs enhances recognition (Kaufmann & Schweinberger, 2012; K. Lee et al., 2000; Mauro & Kubovy, 1992; Rhodes et al., 1987; Schulz et al., 2012), whilst anti-caricaturing (making the stimuli more average) leads to longer reaction times (Rhodes et al., 1987; Schulz et al., 2012) and reduced identification accuracy (K. Lee et al., 2000). Interestingly, caricaturing even improves recognition accuracy in deep convolutional neural networks (M. Q. Hill et al., 2019). Subsequent recognition of veridical faces is enhanced by caricaturing during encoding (Rodríguez et al., 2009), suggesting that exaggerating the features or configuration can help create representations for new faces. Furthermore, adapting to caricatures makes veridical images appear more average (Carbon & Leder, 2005), consistent with the idea that the subset of neurons processing caricatured faces are the same as for their veridical versions. Caricaturing exemplars from the norm also increases the EEG amplitude of the face-selective N170 and N250 ERP responses (Kaufmann & Schweinberger, 2012; Schulz et al., 2012), although other neural responses such as the P200, decreased with distance from average (Schulz et al., 2012), suggesting that some neural processes may encode averageness and typicality.

Studies investigating distance from average on the neural response have adopted a variety of methods making direct comparison difficult (Carlin & Kriegeskorte, 2017; Chang & Tsao, 2017; Davidenko et al., 2012; Leopold et al., 2006; Loffler et al., 2005; McKone et al., 2014; Susilo, McKone, & Edwards, 2010). Chang and Tsao (2017) found near-linear increases with increasing distance through the average in macaques using single unit recordings, as has prior research (Leopold et al., 2006, which included moderately caricatured faces). Likewise, some behavioural work in humans using adaptation has found that the strength of the aftereffect caused by adapting to faces with varying eye and mouth height increased linearly, even outside the range of natural variability (Susilo, McKone, & Edwards, 2010). Other research suggests that the strength of identity aftereffects following adaptation increases linearly, but then is slightly reduced but constant past the ‘naturalness boundary’ (McKone et al., 2014). Results from functional magnetic resonance imaging (fMRI) studies have found saturating responses to stimuli at a certain distance from average (Carlin & Kriegeskorte, 2017; Loffler et al., 2005). The faces in these studies did not extend far past the range of natural plausibility.

Electrical brain stimulation of the fusiform face area (FFA, Kanwisher et al., 1997) produces metamorphosis of viewed faces (Parvizi et al., 2012), suggesting that hyperactivity in the FFA delivers the perception of a caricatured face and thus may represent distance in face space. The perceived change in shape is consistent with suggestions that the FFA is homologous to the area ML in macaques (Tsao et al., 2003, 2008, note the 2003 paper refers to area ML as macaque area pSTS) given that this region shows greater sensitivity to shape over texture (Chang & Tsao, 2017). There is debate, however, over exact homology between human and macaque face processing systems (Rossion & Taubert, 2019; Yovel & Freiwald, 2013).

Hyper-caricatures, images that appear distorted beyond the range of natural appearance, can be generated by extrapolating in face space. In a face space constructed by principal components analysis (PCA), using weights much larger than those corresponding to typical faces shifts the representation further from the mean (see Figure 3.1). This allows the generation of a parametrically controlled set of realistic and hyper-caricatured faces that can be used as stimuli for brain imaging. Specifically, we wanted to explore how the blood-oxygen dependent (BOLD) fMRI signal changes in face-selective cortex, including the FFA, with stimuli at various distances from average in face space and with concomitant changes in perceived naturalness.

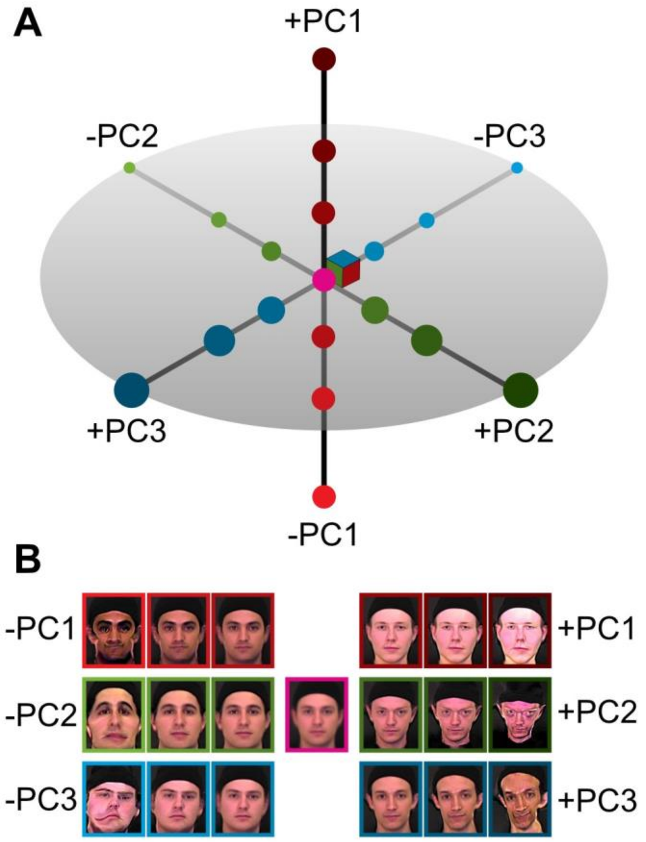


Figure 3.1. Caricaturing in face space.
(**A**) An illustration of the three major axes of a principal component face space constructed from images of male faces. The origin in this space corresponds to the average face. Principal components (PC1 red, PC2 green, PC3 blue) are ordered by variance explained in the underlying data. (**B**) Example images created by modulating each of the principal components independently. Positive and negative deviations from the origin result in opposing changes in reconstructed images, increasingly caricatured with larger distances from the origin (average face, centre).

We hypothesised that there would be an increase in the BOLD response amplitude in the FFA and other face-selective areas with increases in caricature level. To summarise the experimental design, participants first undertook a behavioural session in which they identified the point along different directions in the PCA space where the face stimuli switch from appearing natural to caricatured. The caricature level of stimuli for the fMRI session were then chosen to straddle those perceptual boundaries: some stimuli appeared closer to average and natural, whilst others appeared hyper-caricatured. Stimuli were presented in an event-related design to avoid adaptation to a specific axis (Davidenko et al., 2012; Loffler et al., 2005). Inverted (upside down) stimuli were also presented to identify low-level effects of increased caricaturing (Davidenko et al., 2012). Inverted faces contain the same low-level properties as their upright counterparts, but have been shown to decrease the fMRI response in face-selective areas (James et al., 2013; Nasr & Tootell, 2012; Yovel & Kanwisher, 2004, 2005). We therefore considered that the effect of caricature level might be greater for upright faces than inverted faces.

Our results show that in the right fusiform face area (FFA), *the patterns of fMRI response* were more consistent as caricaturing increased. However, we found no consistent effect of either caricature level or facial inversion on the average fMRI response in the FFA or face-selective regions more widely. Therefore, we also explored the response in object and scene-selective areas. In contrast to face-selective regions, object-selective regions showed an increase in both the consistency of response pattern as well as average fMRI response with increasing caricature level.

3.4 Materials and methods

3.4.1 Participants

Nine healthy, neurologically intact volunteers with normal or corrected-to-normal vision were recruited for this study. Participants were aged between 22 and 36 years old (mean = 27 years, 6 months, SD = 4 years, 1 month). Three were female, six were male. No other demographic details were collected. The sample was a mix of postgraduate research students and staff from the School of Psychology at the University of Nottingham, recruited through a mix of convenience and snowball sampling. All participants gave fully informed consent and were screened for any MRI contraindicators before taking part in the experiment. The study was approved by the School's ethics committee.

3.4.2 Apparatus

The experiment was built in MATLAB version 9.5 (R2018b) using the Psychophysics Toolbox extensions (Psychtoolbox-3 version 3.0.17, Brainard, 1997; Kleiner, 2007; Pelli, 1997). The behavioural experiment was run on a 13" MacBook Pro (1,280 x 800 pixels). Participants responded solely through moving and clicking the mouse. Viewing distance was approximately 60cm. For the MRI experiment, stimuli were presented on a 32", 1,920 × 1,080 pixels BOLDscreen32 (CRS Ltd., Rochester, Kent) with a refresh rate of 120Hz at the back of the bore through a mirror mounted on the head coil. Viewing distance was approximately 120cm.

3.4.3 Stimuli

Stimuli were made using two separate PCA spaces, one derived from 50 images of female faces and another from 50 male faces. The input images were all aligned using the positions of the eyes and then warped to the average of the faces using the Multi-channel Gradient Model (Johnston et al., 1992, 1999), providing shape-free textures as well as the x and y warp information to convert the texture of the face back to the individual's facial shape. The x-y warp fields were appended to the shape-free textures and PCA was performed on these full warp-texture vectors using a procedure described by Nagle and colleagues (2013). The PCA extracts texture and shape covariations and maps these commonalities into an orthogonal space. Face images can be reconstructed by taking the texture for a given position of the PCA space, and spatially displacing the pixels by the distances contained in the corresponding x-y warp fields (see Supplementary Figure 3.1). Reconstructed stimuli were 100 pixels wide by 120 pixels high. In the MRI experiment, the stimuli were feathered into the RGB background around the edges.

To create the stimulus set for the experiment, the first 5 components in each of the PCA spaces were manipulated. The PCA returns eigenvectors of unit length. It also returns values of how the input images load onto each of the components. The components in our space were scaled by 1 standard deviation (SD) of the loadings, such that moving 1 ‘unit’ along a given component reflected a change of 1 standard deviation of the loadings of the input set on that component.

3.4.4 Behavioural task

To establish the caricature levels at which faces turned from natural (physically plausible) to unnatural (physically implausible), we performed a behavioural experiment outside the scanner. This also helped to familiarise participants with the stimuli.

Stimuli scaled the first five components of each gender’s PCA space in both the positive and negative directions (20 possible stimulus directions: 2 gender * 5 PCs * 2 directions), with each unique trial type presented 6 times in a random order - 120 trials in total. Stimuli were presented centrally on a grey background at half the screen height (approximately 11.2^{o} of visual angle).

Using a method of adjustment, participants identified the transition points to unnatural stimuli by moving a mouse. Stimuli were dynamically updated at a caricature level controlled by the horizontal position of the mouse. A red dot on a scale bar served as a visual cue. Before each trial, an animation showed the full range of possible caricaturing for that trial (see Supplementary Figure 3.2a, for demonstration videos see Supplementary Materials). Participants confirmed their choice with a mouse click and the next trials started after a 1000ms inter-stimulus interval. Because some components lead to distortions faster than others, the caricaturing applied to the stimuli was based on some pilot results from 5 independent participants (see Supplementary Table 3.1). Randomly varying the maximal amount of caricaturing on each trial prevented the slider's position being used to indicate the boundary for the given component.

No fixation cross was presented so participants could freely explore the faces, and there was no time limit. Breaks were provided every 40 trials. On average participants took approximately 30 minutes to complete the experiment.

For each participant, the average naturalness boundary for each component was calculated by taking the mean transition point across the 6 repetitions. The value of this position on the scale translated to the number of standard deviations (in terms of the loadings of the input set onto the PCA space) from the origin of the space. The results of the first 7 participants were used to scale the stimuli for the MRI experiment (see Supplementary Figure 3.3 and Supplementary Table 3.2). Results of all participants can be seen in Supplementary Table 3.3.

3.4.5 MRI study

3.4.5.1 Localiser and caricature scans

The fMRI study consisted of two sets of scans. To find cortical regions responding to various categories of stimuli, we ran a standard functional localiser experiment using a randomised block design. We also ran a set of event-related scans in which individual images of test stimuli were presented (“caricature scans”).

In the functional localiser, images of faces, scenes and objects were presented in a block design. Each block consisted of 8 images from one category. Face stimuli included photographs of 24 different identities (12 male, 12 female) taken at frontal pose, and 45^{o} rotated in yaw in either direction. Not all views of each identity were presented. Images of scenes included both natural and manmade scenes, including pictures of buildings, both from the inside and outside. Objects included both manmade and natural objects. Faces and objects were presented on greyscale masks to occupy the same space as the scene stimuli (see Supplementary Figure 3.4). All stimuli were presented centrally and extended to approximately +/- 8^{o} of visual angle. Each stimulus was presented for 1s with no ISI, with 8 s between blocks giving an 8s ON, 8s OFF sequence. The experiment began and ended with 8s OFF. During the localiser a simple attention task was used: a black fixation cross was presented centrally throughout which 130 times within a scan turned red for 50 ms and participants had to respond by pressing any button on the button box. Any response within 1.5s was classed as a hit. Each run of the functional localiser took 6min and 32s.

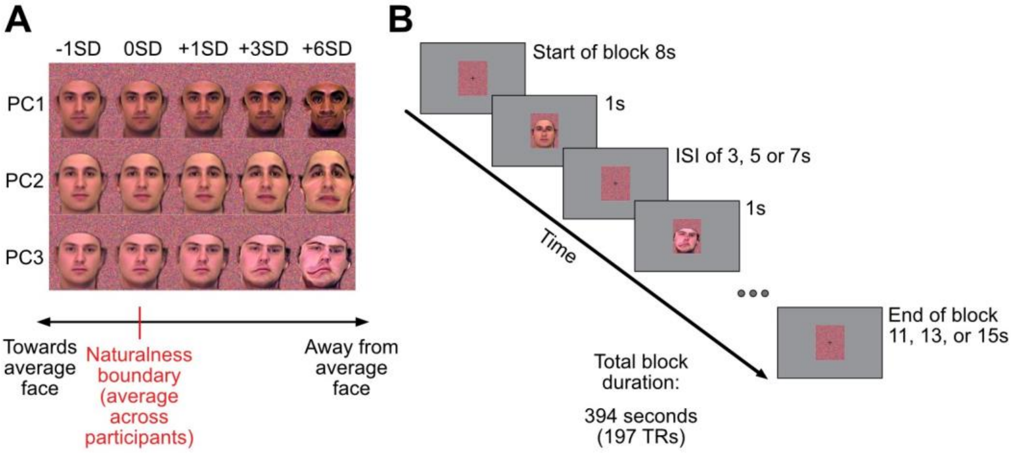


Figure 3.2. Caricatured stimuli used and outline of the fMRI experiment.

(**A**) Example images corresponding to the five scaling levels used in the MRI experiment. First column of images: stimuli that are 1 standard deviation (SD) closer to the average face. Second column: group-averaged naturalness boundary for a given component (0SD). Other columns: images corresponding to +1, + 3 and +6SD away from average. Stimuli were presented on colour masks (Gaussian noise on each R, G, B channel, with mean and standard deviation derived from the face stimuli). (**B**) Timings for trials in the event-related caricature scan. The experiment started and ended with 8 s of fixation (dynamically changing coloured masks only). Stimuli were presented for 1s, followed by a variable inter-stimulus interval of 3, 5 or 7s. To control for attention, participants had to report colour changes of the fixation cross (black to red), which occurred randomly 42 times within each run. The background colour mask changed dynamically every second.

During the caricature scan participants were presented with stimuli created by modulating the first three components of the male PCA space from the behavioural study. Using the averaged naturalness boundaries from the behavioural experiment, participants were shown faces that corresponded to the mean (across participants) naturalness boundary (0SDs), one SD (across participant responses) closer to the average face (-1SD), or one, three or six SDs further away from the average (see Figure 3.2a).

In each run, participants were presented with 5 caricature levels for each component (the average naturalness boundary plus -1, 0, +1, +3, and +6SDs). Picture plane inverted images of the most (+6SDs) and least caricatured (-1SD) face stimuli were also presented. This provided 21 unique stimuli per run (15 upright and 6 inverted) which were repeated 3 times each in a run. The experiment started with 8s of rest. Subsequently stimuli were presented for 1s with a variable ISI of either 3, 5, or 7s, with equal numbers of each ISI duration across each run. Trial timings can be seen in Figure 3.2b. The order of stimuli and ISI durations was pseudorandomised across runs. To ensure all runs were the same duration, the final stimulus of each run was always followed by the remaining ISI and a further 8 s of rest (minimum of 11s in total) to allow for the lag in the haemodynamic response. Each run lasted 6min and 34s. As in the localiser scan, participants responded when the centrally presented fixation cross turned red. This occurred 42 times during the run.

3.4.5.2 Data acquisition

For the localiser, functional data were acquired across 2 block-design runs, each lasting 392s (196 volumes), one at the start of the scanning session and one at the end. Caricature scans were acquired across 3 event-related runs (4 for one participant), each lasting 394s (197 volumes).

Data were acquired on a 3T MRI scanner (Phillips Achieva) at the Sir Peter Mansfield Imaging Centre at the University of Nottingham using a standard 32 channel head coil. Functional (BOLD) images were acquired with 2D gradient echo EPI sequence (multiband 2, SENSE r = 1). Parameters were TR/TE 2000ms/32ms, FA 77º. There were 34 axial slices; voxel size was 2.4 x 2.4 x 3mm, 80 x 80 voxels per slice. High-resolution T1 MPRAGE structural images were obtained with the following parameters: TR/TE 8.1 ms/3.7 ms, 1 mm isotropic voxels, 256 x 256 voxels, FOV = 256 x 256 mm, 160 sagittal slices.

3.4.5.3 Data analysis

We used a combination of tools to analyse fMRI data: mrTools (Gardner et al., 2018) and custom MATLAB code, as well as FreeSurfer (Fischl, 2012) for cortical segmentation and anatomically defined regions of interest and FSL (Jenkinson et al., 2012) for spatial smoothing and mask dilation. Analyses were performed in individual participant space.

3.4.5.4 Anatomically restricting the analyses

We focused our analysis on the occipito-temporal cortex, bilaterally, including the FFA, the OFA (occipital face area, Halgren et al., 1999; Puce et al., 1996) and pSTS (posterior superior temporal sulcus, Morris et al., 1996, 1998). We defined larger anatomical ROIs from FreeSurfer parcellations to span the majority of the occipito-temporal cortex, spanning both hemispheres (combining ‘lateraloccipital’, ‘fusiform’, ‘inferiortemporal’, ‘middletemporal’, ‘superiortemporal’, ‘bankssts’, ‘supramarginal’ and ‘inferiorparietal’ ROIs from the Deskian/Killiany atlas). ROIs were created by converting the parcellation labels into volumetric masks (FreeSurfer: *mri_annotaion2label* and *mri_label2vol*) and dilated using a single pass of a 3-voxel box kernel to fill any holes (*fslmaths*).

3.4.5.5 Pre-processing

The caricature and localiser scans were first motion corrected within and between scans in mrTools (Gardner et al., 2018) using the mean volume of the second caricature scan (mid-point of the scanning session) as a reference frame. Motion correction used linear interpolation and drift correction was applied. The motion corrected functional runs were then spatially aligned to the participants’ anatomical scans. The localiser data was spatially smoothed (3D Gaussian, FWHM 5mm). For the caricature data, voxelwise data was extracted from the face, object, and scene-selective ROIs. For the univariate analysis and multivariate pattern analysis on the data no spatial smoothing was applied.

For both the localiser and caricature scans, data were converted to percentage signal change by subtracting the mean intensity for each voxel across the scan, and dividing by the mean ([*x*-mean]/mean), temporally high-pass filtered (cut-off 0.01 Hz) and, for the univariate analysis, concatenated over scans, taking care to keep track to the transition points between scans. This allowed for the GLM analysis to be reframed in block matrices, requiring only one GLM per set of localiser and caricature scans.

3.4.5.6 Defining the FFA and face-selective, object-selective, and scene-selective voxels

To define participant-specific functional ROIs, we used a GLM approach and restricted the analysis to the anatomical ROI described above. Analyses were performed in individual scan space. The 3 explanatory variables (EVs) were faces, objects and scenes, specified by 8s ON boxcar regressors convolved with a double gamma haemodynamic response function (HRF). To define face-selective areas responses to face blocks were compared to blocks of objects and scenes (faces > objects + scenes). Voxels that responded significantly more to faces over objects and scenes were defined as face-selective. Corresponding contrasts then defined object-selective (objects > faces + scenes) and scene-selective areas (scenes > faces + objects). We used family-wise error (FWE) correction to account for multiple comparisons.

The functional ROIs were then defined on flat map representations of the corresponding statistical maps. A cluster corresponding to the FFA was present in each participant bilaterally (for details see Supplementary Table 3.4), however, in some participants, the boundaries were less clear, and even with family-wise error correction extended further along the fusiform gyrus and even into the neighbouring sulcus. In these cases, the FFA was defined as one contiguous cluster within a region restricted anatomically to the fusiform gyrus (from a FreeSurfer parcellation, FFA definition in each participant can be seen in Supplementary Figure 3.5). The pattern of response elsewhere however was more variable. Therefore, rather than trying to identify spatially consistent ROIs across participants, we simply allocated voxels to the 3 categories ‘face-selective’, ‘object-selective’ and ‘scene-selective’ based on the contrasts above. Face-selective voxels included the FFA. Voxels that responded significantly to more than one contrast were removed, such that each ROI only contained voxels that exclusively appeared for that contrast. Functional ROIs from one participant can be seen in Figure 3.3b-d.

3.4.5.7 Univariate analysis

To assess the effect of caricature level in the FFA, and face-, object- and scene-selective areas, we first used a deconvolution analysis (e.g., see Besle et al., 2013; Gardner et al., 2005). This provided an estimate of the event-related BOLD response for each of the 7 stimulus types (5 caricature levels, upright images; 2 inverted images). From these event-related responses (see Figure 3.3e), we calculated an index of the response amplitude of the first 5 TRs after stimulus onset, by first normalising to the level at stimulus onset (the first TR) and then obtaining the mean signed deviation (MSD) across the subsequent four TRs.

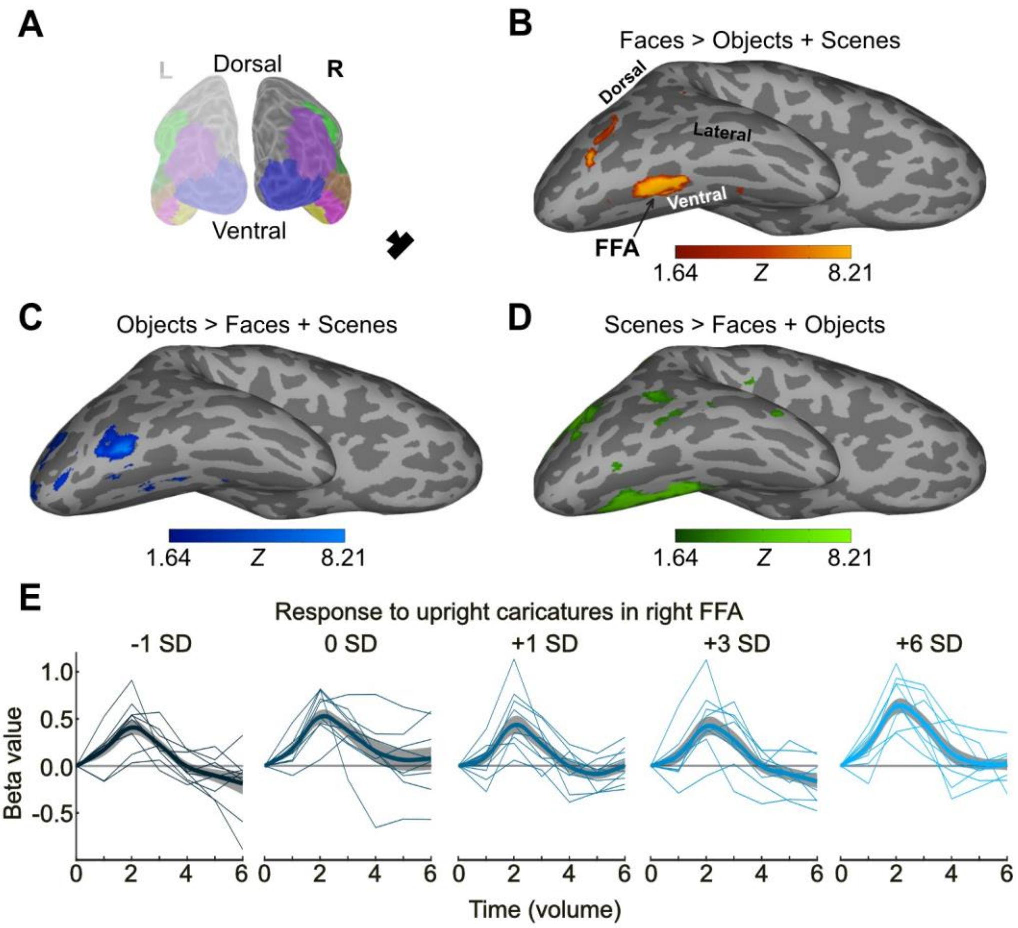


Figure caption on next page

Figure 3.3. Defining face, object, and scene-selective regions and event-related fMRI responses in right FFA.
(**A**) Posterior view of left and right inflated cortical hemispheres. Camera symbol indicates the view in panels B-D. Light grey, gyri; dark grey sulci. Regions in colour, Freesurfer parcellations used to form the bilateral occipitotemporal ROI, including lateral occipital (blue), fusiform gyrus (yellow), inferior temporal (pink), middle temporal (brown), bank of the superior temporal sulcus (dark green), inferior parietal (purple) and supramarginal gyrus (light green). (**B**) Face-selective regions in one participant based on the contrast faces > objects + scenes from the localiser scans (FFA, fusiform face area). Object-selective (**C**) and scene-selective (**D**) voxels defined using the contrasts objects > faces + scenes, and scenes > faces + objects, respectively. The colour bars in B-D show the Z-statistic for the contrast, thresholded at Z > 1.64 (corresponding to *p* < .05, with FWE correction). Maps show the voxels exclusively defined by these contrasts, with any overlap removed. (**E**) Response amplitude in the right FFA across participants from stimulus onset as a function of time for the five different levels of caricaturing (upright only) from the deconvolution analysis. Y values show the beta-coefficients from the deconvolution, normalized to t=0. Thin lines show the average timeseries for each participant. Thick lines show the group average, smoothed over time (for display purposes only). Shaded areas show ±1 SEM across participants. Solid grey line shows Y=0. Colour represents the caricature level.

3.4.5.8 Multivariate pattern analysis

To look at patterns of response across the regions of interest, we also performed a correlation-based multivariate pattern analysis (MVPA). We compared the correlations in response patterns (beta values) between all 5 caricature levels of upright stimuli. The analysis was performed on the left and right FFA, left and right face-selective cortex, and left and right object-selective cortex.

The β values were obtained for each caricature scan repeat separately using a GLM similar to that described above, but assuming a canonical haemodynamic response function (double gamma). There were 5 explanatory variables (one for each upright caricature condition). The two additional conditions (inverted stimuli) were included as nuisance regressors. For each region of interest in the analysis, we then calculated the correlations of the β coefficient maps across regressors (avoiding within-scan comparisons). We then applied Fisher's transform to convert from correlation, r, to Z and averaged these Z-values across scans for each participant separately.

3.5 Results

The average fMRI response in the FFA, as well as face-selective voxels overall, did not show a consistent change with either caricature level or inversion, as assessed by univariate analysis and ANOVA. Interestingly, however, in the right fusiform face area (FFA), we found that the patterns of fMRI response were more consistent as caricaturing increased as assessed by multivariate pattern analysis (MVPA). In contrast, object-selective regions showed an increase in both average fMRI response with increasing caricature level (univariate analysis), and the consistency of response pattern (MVPA).

3.5.1 Univariate analyses

To assess the effects of caricature level and inversion in the FFA we performed two separate within-subjects ANOVAs, one to assess the effect of caricature level and orientation using the least and most caricatured faces, and one to assess the effect of caricature level using all 5 caricature levels of upright stimuli. The first was a 2 x 2 x 2 ANOVA with hemisphere (left, right), stimulus orientation (upright, inverted) and caricature level (-1SD, +6SD), the second a 2 x 5 ANOVA with hemisphere (left, right) and caricature level (all 5 levels of upright caricature). ANOVAs were performed using IBM SPSS Statistics version 25.

To investigate the response amplitudes in the face-, object- and scene-selective voxels we performed the same two ANOVAs as for the FFA but including ROI as an additional independent variable with 3 levels (face-selective, object-selective, and scene-selective).

3.5.1.1 Caricature level in the FFA.

The event-related response profiles showed a clear trial-locked response to the 5 caricature levels across all regions. Figure 3.3e shows the average deconvolution timeseries for the right FFA across subjects (thick lines), as well as traces for individual participants (thin lines).

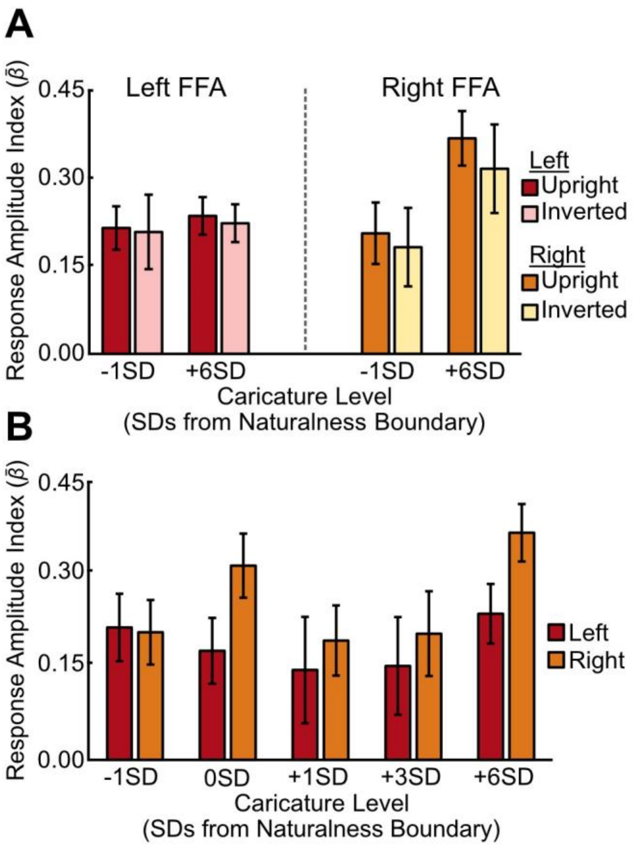


Figure 3.4. Response amplitudes in the right FFA for different stimulus conditions.

(**A**) Average response amplitude in the left (red/pink) and right (orange/yellow) FFA to the most (+6SD) and least (-1SD) caricatured faces in both the upright and inverted conditions. (**B**) Average response amplitudes in the left (red) and right (orange) FFA to each of the five caricature conditions for the upright stimuli only. Bars are grouped according to deviations from the average naturalness boundary (-1SD, closer to the average face; +6SD is highly caricatured). Y axes show the response amplitude index, measured by offsetting the *β* coefficients from the deconvolution analysis by t_0, and averaging $t_{1\text{-}4}$. Error bars show ±1SEM across participants.

When assessing caricature level (-1SD, +6SD), including both upright and inverted stimuli, there was no main effect of caricature level ($F(1,8) = 3.08$, $p = .117$, $\eta_p^2 = .28$), but there was a significant interaction between hemisphere and caricature level ($F(1,8) = 5.86$, $p = .042$,

η_p^2 = .42). The interaction was driven by a stronger increase in the response amplitude in the right FFA than the left FFA (t(8) = 2.42, p = .042) to an increase in caricature level (Figure 3.4a), although the effect of caricature level in the right FFA was marginal (F(1,8) = 5.18, p = .052, η_p^2 = .39).

We found no interaction between hemisphere and caricature level when we assessed all 5 levels of the upright stimuli (F(4,32) = 1.59, p = .200, η_p^2 = .17, Figure 3.4b), nor a main effect of caricature level (F(4,32) = 1.99, p = .119, η_p^2 = .20).

3.5.1.2 Caricature level in face, object and scene-selective regions.

We also compared responses across face-selective regions more generally, as well as in object- and scene-selective regions (see Supplementary Table 3.5 for details). The data are shown in Figure 3.5.

We found no significant main effect of caricature level when assessing the effect of caricature level (extremes) and inversion (Figure 3.5a), but there was a significant interaction with ROI (F(2,16) = 6.08, p = .011, η_p^2 = .43). This interaction showed the effect of caricature level was only present in the object-selective cortex, with the object-selective cortex increasing in response amplitude with an increase in caricature level (t(8) = 2.49, p = .038).

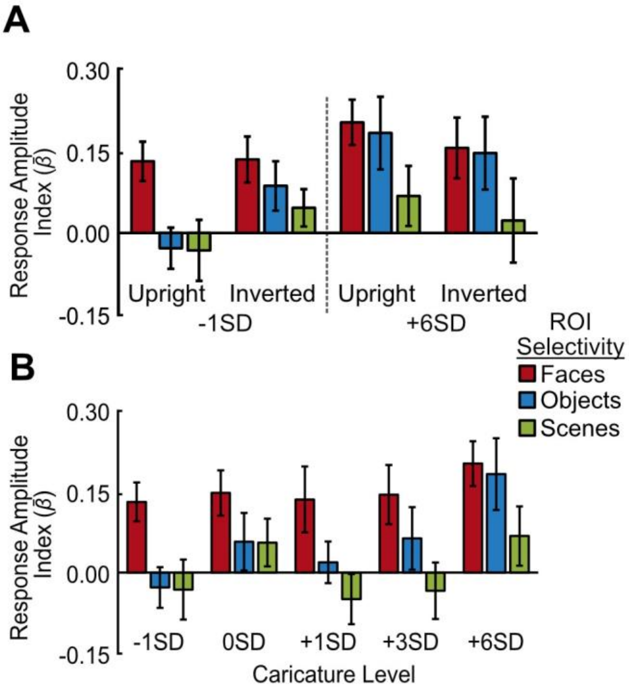


Figure 3.5. Response amplitudes to caricatured faces in face-, object- and scene-selective voxels.
(**A**) Average response amplitudes for least (-1SD) and most (+6SD) caricatured stimuli. Data are grouped by image orientation (upright, inverted) and region of interest (face-, object-, and scene-selective voxels across hemispheres). (**B**) Average response amplitude to the 5 levels of caricatured, upright faces in face-selective (red), object-selective (blue) and scene-selective (green) voxels. As there was no interaction with hemisphere the response amplitudes are averaged across hemispheres. Y axes show the response amplitude index, measured by offsetting the *β* coefficients from the deconvolution analysis by t_0, and averaging $t_{1\text{-}4}$. Error bars show ±1SEM of the between-subjects variance.

When assessing all 5 levels of upright caricature, there was a main effect of caricature level ($F(4,32) = 3.17$, $p = .027$, $\eta_p^2 = .28$) driven by a general increase in response amplitude as a function of caricature level, which was particularly prominent for highly caricatured (+6SD) faces. The ANOVA showed there to be a positive linear trend between response amplitude and caricature level (*F*(1,8) = 16.83, *p* = .003, η_p^2 = .68). Highly caricatured faces (+6SD) elicited a stronger response than -1SD (*t*(8) = 4.83, *p* = .001), 0SD (*t*(8) = 2.47, *p* = .039) and +1SD (*t*(8) = 3.22, *p* = .012) caricatures, although only the first of these survived Bonferroni-correction (*a* = .005).

Although the interaction did not reach significance (*F*(8,64) = 1.96, *p* = .066, η_p^2 = .20), the overall effect was primarily driven by object-selective regions. The data in Figure 3.5b shows a constant response across caricature level in face-selective and scene-selective areas, but an increase in the response amplitude with increasing caricature level in the object-selective areas. To support this, separate ANOVAs for each ROI revealed that in the face-selective and scene-selective regions there was no effect of hemisphere, caricature, nor any interaction. In the object-selective regions there was no main effect of hemisphere nor interaction, but there was a significant effect of the caricature condition (*F*(4,32) = 4.76, *p* = .004, η_p^2 = .37) paired with a positive linear effect (*F*(1,8) = 26.69, *p* = .001, η_p^2 = .77). Highly caricatured faces (+6SD) again elicited a stronger response over -1SD (*t*(8) = 5.31, *p* = .001), 0SD (*t*(8) = 2.82, *p* = .023) and +1SD faces (*t*(8) = 3.97, *p* = .004). The difference between +6SD and 0SD was not significant when correcting for multiple comparisons (*a* = .005). Interestingly faces on the naturalness boundary (0SD) also elicited a greater response than the most average (-1SD) faces (*t*(8) = 2.58, *p* = .032).

3.5.1.3 Effects of ROI, orientation, and hemisphere.

We found that there was a decrease in response amplitude from face, to object, to scene-selective cortex in response to our face stimuli. Main effects of ROI were significant when assessing the response to upright and inverted, -1SD and +6SD caricatured stimuli ($F(1.27,10.12) = 9.63$, $p = .008$, $\eta_p^2 = .55$, Greenhouse-Geisser correction applied) and when assessing all 5 levels of upright stimuli ($F(2,16) = 9.95$, $p = .002$, $\eta_p^2 = .55$). All pairwise comparisons were significant prior to correction (all $p < .046$) with the difference between face and scene-selective regions surviving correction ($a = .017$) in both analyses (both $p < .012$).

We found no main effects of, nor interactions with, orientation in any of our ROIs, and there were also no significant main effects of hemisphere. Generally, there was a greater response amplitude in the right hemisphere ROIs, which was most notable in the FFA when assessing the response to all five upright caricature levels ($F(1,8) = 4.96$, $p = .057$, $\eta_p^2 = .38$).

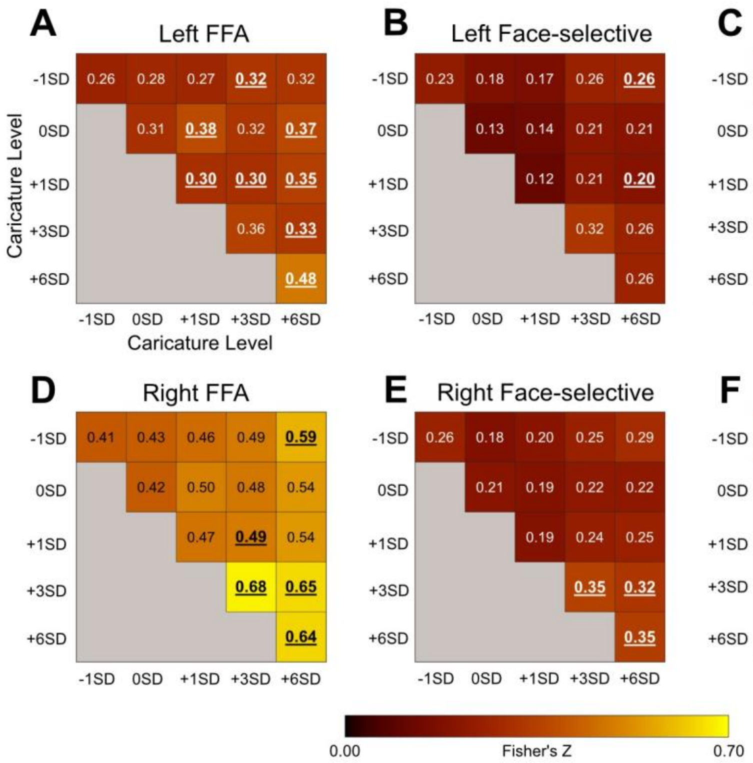


Figure 3.6. MVPA results.

Plots showing the Fisher’s Z for the correlation coefficients between *β* maps corresponding to the 5 upright levels of caricature from the MVPA analysis in the left FFA (**A**), the left face-selective voxels (**B**), left object-selective voxels (**C**), the right FFA (**D**), right face-selective voxels (**E**), and the right object-selective voxels (**F**). The diagonal reflects the average correlations between the response patterns to stimuli of the same caricature level, while the off-diagonal reflects correlations between different caricature levels. Only between-scan correlations were assessed. Values in bold and underlined were significantly greater than 0 at a group level, measured using one-sample t-tests (Bonferroni-corrected *a* = .003). Font colour for display purposes only.

3.5.2 Multivariate pattern analysis

The results of the correlation analysis can be seen in Figure 3.6. In each ROI, we tested whether there was a significant positive correlation in the response patterns for each pair of caricature levels. Significance was assessed using one-sample t-tests to test if the group-level Z-value was significantly greater than 0 (Bonferroni-corrected *a* = .003) and is indicated by bold, underlined values in Figure 6. We then assessed how the response patterns varied as a function of caricature level using a one-way within-subjects ANOVA with the 5 levels of ‘same’ caricature correlations (i.e., the diagonals in Figure 3.6) as the independent variable.

3.5.2.1 Caricature level in the FFA

In the right FFA, the correlation coefficient (converted to Fisher’s Z) increased as a function of caricature level (Figure 3.6d), supported by a significant positive linear trend (*F*(1,8) = 7. 60, *p* = .025, η_p^2= .49), indicating increasing consistency in the patterns of responses between stimulus categories including highly caricatured faces. In the left FFA, many of the correlations were significant at a group level, but the overall increase with caricature level, as seen in the right FFA, was not.

3.5.2.2 Caricature level in face and object-selective cortex

When looking at the consistency of response patterns in face and object-selective regions more broadly, we found that only object-selective regions bilaterally showed an increase in consistency with caricature level. Right face-selective regions were sensitive to caricature level, but the change in response pattern was less clear.

In the left face-selective regions there was no significant effect of caricature level on the correlations. In the right face-selective cortex there was a significant main effect of caricature level (*F*(4,32) = 5.06,

p = .003, η_p^2 = .39) however the response profile was less clear than in the right FFA.

In both the left and the right object-selective regions there was a positive linear trend in the correlation to same caricature level trials as a function of caricature level (left: $F(1,8)$ = 21.91, p = .002, η_p^2 = .73; right: $F(1,8)$ = 10.09, p = .013, η_p^2 = .56). At the group level however, no Z-values were greater than 0 in the left object-selective cortex. In the right hemisphere only correlations between +3SD and +3SD stimuli and +3SD and +6SD stimuli were significant.

3.6 Discussion

We investigated the effect of caricaturing on the fMRI response in visual areas defined by preference to faces, objects, and scenes. Based on evidence of ramp coding in single cell recordings in macaques (Chang & Tsao, 2017; Leopold et al., 2006) to face stimuli of increasing distance from the mean face in the neuron's preferred direction of change, we reasoned that there may be an increase in the response amplitude of the FFA with increasing caricature level, even when faces appeared heavily distorted.

Surprisingly, we found no clear change in the average response amplitude in the FFA, or face-selective cortex more broadly, with increasing caricature level. In contrast, we found an increase in response in object-selective cortex, particularly for highly caricatured faces. There was no significant change in response in scene-selective areas.

An increase in the consistency of the response pattern in object-selective cortex was also observed with increasing caricature level, measured using MVPA. Caricaturing therefore both enhanced the average response, and the consistency in which the stimuli were processed within object-selective cortex. How or why caricatured faces activate object-selective cortex is unclear.

The results seen in object-selective regions may result from changes to low-level or even mid-level properties that vary with caricature level, rather than a response to caricatured faces *per se* or the assignment of hyper-caricatured faces to a separate object class other than faces. Higher-level visual regions, including the FFA (Weibert et al., 2018), are sensitive to the lower-level image properties that are characteristic of different categories of objects (see Andrews et al., 2015). Caricaturing may have therefore emphasised particular low or mid-level properties that object-selective neurons are tuned to, such as certain shapes or curvatures that distinguish animate faces, bodies and animals from inanimate objects (Yetter et al., 2021; Yue et al., 2020; Zachariou et al., 2018) or changes in bilateral symmetry (Bona et al., 2015). The areas defined as object-selective responded *more* to objects than faces and scenes despite many voxels responding to all three categories (see Supplementary Figure 3.6) so the changes with caricature level may have generated stimulus properties that are more characteristic of generic objects than faces. The changes in our stimuli, including changes in texture, colour (Lafer-Sousa et al., 2016), shape, curvature (Yetter et al., 2021; Yue et al., 2020) and external contours, as well as higher-level changes, may have caused a shift in object-space (Bao et al., 2020).

Regardless of the exact mechanism for why or how caricatures activate object-selective cortex, it is evident that object-selective cortex is sensitive to caricature level, raising the possibility of its involvement in the perceptual evaluation of faces. To our knowledge, these are the first findings that show that caricatured faces elicit increased responses in regions typically involved in processing objects. These findings can potentially have important implications for understanding how we might form impressions from or recognise more distinctive faces. Although the faces in our experiment were artificially caricatured, faces in the real world can be naturally distinctive too, for example a number of (often genetic) disorders give rise to naturally distinctive faces (Babovic-Vuksanovic et al., 2012; Dolci et al., 2021; Gad et al., 2008). Our findings therefore raise a number of questions as to whether, and if so how, object-selective cortex contributes to our social evaluation of faces.

Returning to face-selective cortex, we initially found evidence that our most caricatured faces elicited a stronger response in the right FFA (but not left) compared to our least caricatured faces when we included both upright and inverted stimuli. This interaction between hemispheres is potentially consistent with the idea of a greater involvement of the right FFA in face perception, for example, electrical brain stimulation only impacts perception of faces when applied to the right FFA and not the left (Rangarajan et al., 2014).

We found no evidence of an effect of caricature level in face-selective cortex however when we assessed for a graded change in response amplitude across the complete range of caricature levels; there was no effect of hemisphere, caricature level nor an interaction, in either the FFA, specifically, or face-selective regions more widely. This may reflect a plateau in the BOLD response (Carlin & Kriegeskorte, 2017; Loffler et al., 2005; McKone et al., 2014). Since most of the stimuli were ‘caricatured’ to some degree the results could reflect response saturation; even the least caricatured stimuli could be identified as a particular individual. Alternatively, since Chang and Tsao (2017) report ramp-like tuned cells which increase their firing along an axis passing through the mean face, increasing caricature level may increase the firing of some cells while reducing the firing of others, leading to no net increase in the response across the population within a voxel.

Interestingly, the multivariate pattern of response in the right FFA became more consistent with increasing caricature level. This suggests a pattern of systematic increases and decreases in response rate across a population of cells. For the right face-selective regions more broadly, the change in spatial consistency was less clear, with slightly increased consistency for more caricatured faces, but decreased spatial consistency for intermediate caricatures. In the left hemisphere there was no effect of caricature level, consistent with the functional differences in the left and right FFA. The increase in spatial consistency in the absence of an average increase in the fMRI response in the right FFA is particularly interesting since stimuli at the same caricature level could vary substantially in terms of their low-level properties, given that different PCA components were modulated. Despite these low-level differences, which became more pronounced as caricaturing increased, the response pattern became significantly more consistent. This indicates that the increase in the overall consistency of the pattern was maintained regardless of any variation in the individual patterns themselves. Likewise, the general increase in consistency appeared to hold regardless of whether we compared the response patterns between the same caricature level, or different caricature levels. This suggests that different levels of caricature are processed by the same set of voxels, and is consistent with the idea that voxel-wise responses scale with varying distances from the average in a face space.

Our analysis of inversion showed no effect of orientation nor interaction with it. The lack of effect of orientation is at odds with some prior research showing an inversion effect in FFA (James et al., 2013; Nasr & Tootell, 2012; Yovel & Kanwisher, 2004, 2005), but is in line with other findings in literature suggesting that the inversion effect is weak (Gilaie-Dotan et al., 2010) or even absent (e.g., see Aguirre et al., 1999; Epstein et al., 2006; Haxby et al., 1999). The initial evidence of an effect of caricature level alongside a lack of inversion effect potentially suggests that the FFA is sensitive to changes in low-level properties, consistent with prior evidence (Weibert et al., 2018) and that this may not be specific to upright faces.

To conclude, in the FFA and face-selective areas more generally, we found no substantive effect of caricature level on the average response amplitude, although we did find evidence that the right FFA is sensitive to caricature level using MVPA, with the consistency of the response pattern increasing with caricature level. In contrast, we found a significant increase in both the response pattern consistency and the average response amplitude in object-selective cortex to increasing caricature level. This suggests that caricatured faces might recruit cortex typically defined as object-selective, potentially because they share more low-level features with objects. This may have implications for understanding how distinctive faces might be processed, both in terms of recognition and forming impressions.

3.7 Supplementary materials

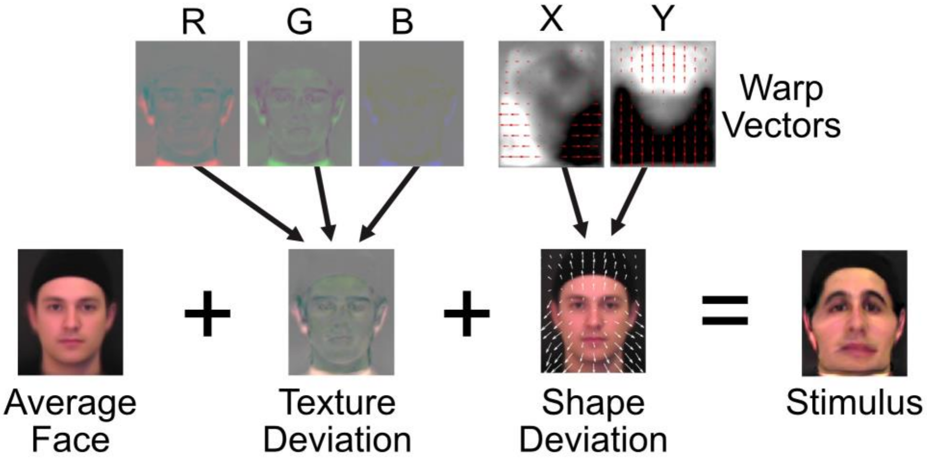
3.7.1 Supplementary videos

Supplementary videos can be found using the link below:

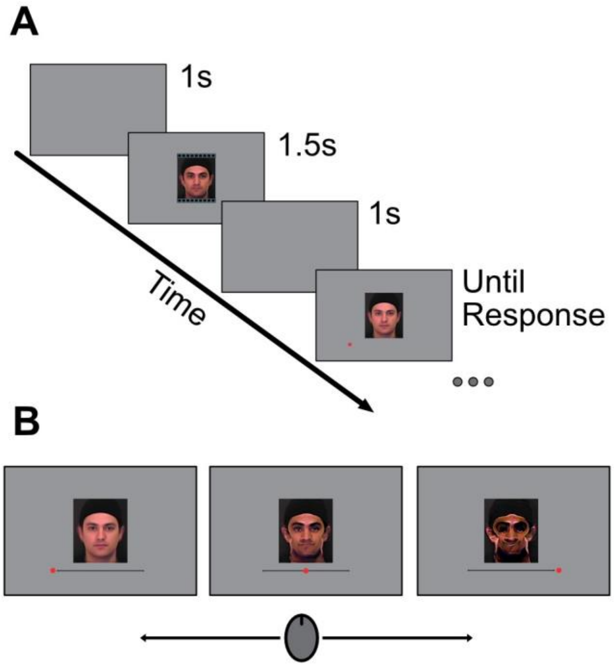
https://www.dropbox.com/scl/fo/kaqogna18e0a5vira9syo/h?rlkey=m0k5mrhvglk69rpjdy92lk5ha&dl=0

Within the Supplementary Materials folder, navigate to 'Supplementary_Videos/Chapter_3_fMRI_study'. This folder contains videos of the first three components of the male space. There is also a Word document to provide information about the videos.

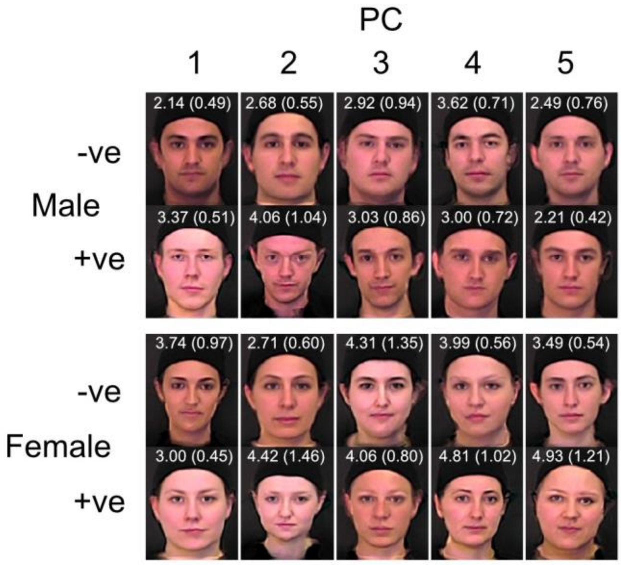
3.7.2 Supplementary figures



Supplementary Figure 3.1. Creating stimuli by adding deviation vectors to the origin of face space.
Stimuli were created by adding texture and shape deviations to the image corresponding to the origin of the PCA-based face space. The texture deviation stores how much of each RGB channel is to be added to the average face for each pixel. Note that for easier visualisation the RGB changes have been exaggerated here. The x-y warp fields contain the horizontal (x) and vertical (y) pixel displacements necessary to distort the average face shape. Lighter areas show leftward and upward displacements, while darker areas show rightward and downward displacements.

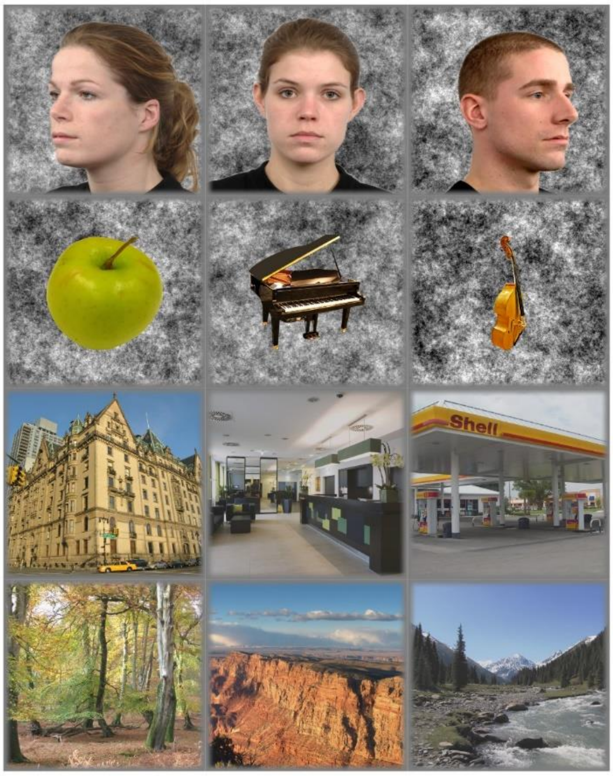


Supplementary Figure 3.2. Depiction of behavioural experiment. Trial timings (**A**) showing an inter-trial interval of 1 second, followed by the demonstration video lasting 1.5s, followed by a 1s inter-stimulus interval before the appearance of the manipulable stimulus. The demonstration video starts with a close to average face that becomes heavily caricatured and then returns to average. (**B**) The manipulable stimulus is presented until response and can be manipulated by the participant to appear more or less average by moving the mouse left to right respectively. When the participant has found the boundary between natural/physically plausible and unnatural/physically implausible the participant responds by clicking either key on the mouse.

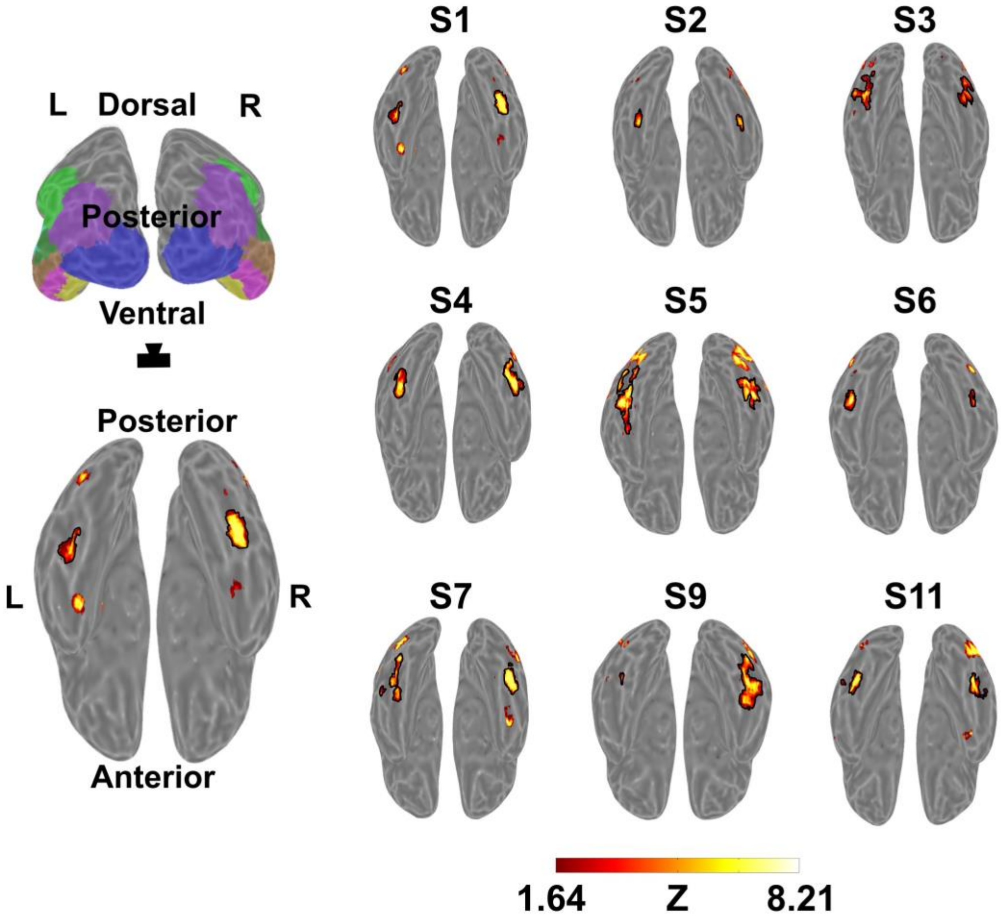


Supplementary Figure 3.3. Group average naturalness boundaries.

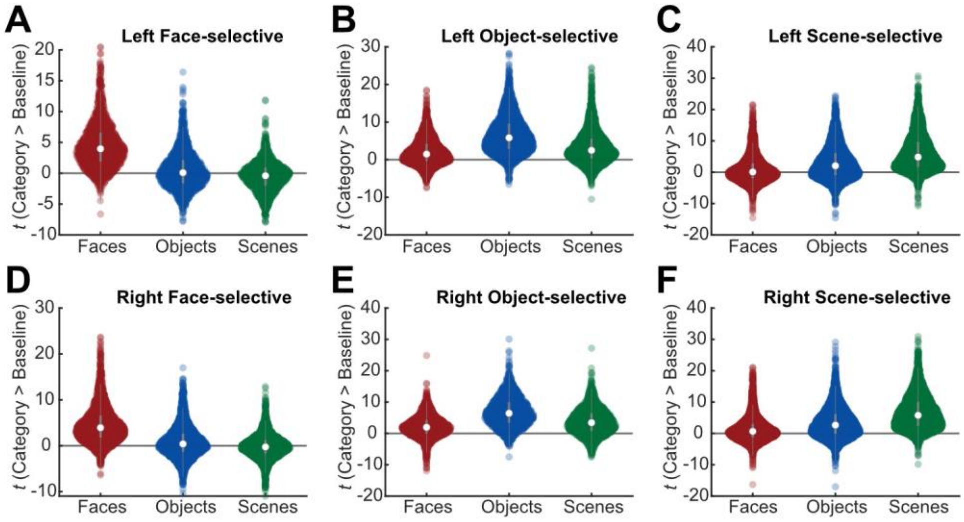
Stimuli representing the average transition point for the first seven participants between natural, physically plausible and unnatural, physically implausible. Rows, two directions (sign) in PCA space for each gender. Columns, five principal components scaled according to the transition points. Numbers, average transition point in terms of the number of standard deviations in PCA space (with respect to the input data) along the given component. Numbers in parentheses, the between-subjects standard deviation from the behavioural results.



Supplementary Figure 3.4. Example stimuli for the localiser scan. Example stimuli for the localiser scan showing faces, objects, manmade scenes and natural scenes. Manmade and natural scenes formed one block type encompassing scenes as a whole.



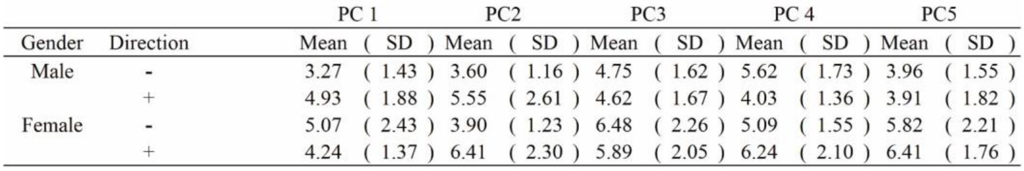
Supplementary Figure 3.5. Defining the FFA in all participants. The two leftmost images depict the current view, with images taken from underneath the ventral surface. The left and right image of each pair show the left and right hemispheres respectively. The images S1-S9 show the FFA definition in each participant (the region surrounded by the black border). The statistical maps show the face-selective regions defined by the contrast faces > objects + scenes that survived FWE correction. Maps show the z-values of the contrasts.



Supplementary Figure 3.6. Violin plots showing the distribution of responses to faces, objects and scenes.

Violin plots to show the distribution of responses to faces (red), objects (blue) and scenes (green) compared to baseline (t-value of the contrast). (**A-C**) The responses in the face-selective, object-selective and scene-selective voxels of the left temporal cortex. (**D-F**) The responses in the right temporal cortex. Above the zero-line show greater responses to the stimuli than baseline, beneath show inhibited responses to the stimuli compared to baseline. ROIs were defined by contrasting the response between different stimulus classes, e.g., face-selective was defined as voxels responding significantly more to faces than objects and scenes. Voxels that responded significantly in more than one of these contrasts were excluded.

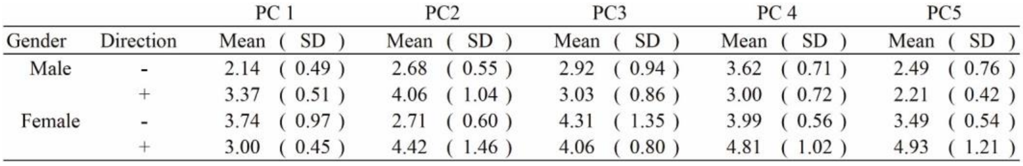
3.7.3 Supplementary tables



		PC 1		PC2		PC3		PC 4		PC5	
Gender	Direction	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Male	-	3.27	(1.43)	3.60	(1.16)	4.75	(1.62)	5.62	(1.73)	3.96	(1.55)
	+	4.93	(1.88)	5.55	(2.61)	4.62	(1.67)	4.03	(1.36)	3.91	(1.82)
Female	-	5.07	(2.43)	3.90	(1.23)	6.48	(2.26)	5.09	(1.55)	5.82	(2.21)
	+	4.24	(1.37)	6.41	(2.30)	5.89	(2.05)	6.24	(2.10)	6.41	(1.76)

Supplementary Table 3.1. Pilot data: Average naturalness boundaries across each stimulus type (between subject SDs).

Pilot data for the behavioural study, showing the group (n=5) average naturalness boundary (between-subjects SD) for the 5 PC components for male and female spaces, in both the positive and negative directions. These values were rounded to 0.d.p and used to scale the stimuli in the main experiment. Stimuli in the main experiment started at 0.3x the boundaries in the table, and scaled to a random value between 1.3x and 1.6x the boundaries.



		PC 1		PC2		PC3		PC 4		PC5	
Gender	Direction	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Male	-	2.14	(0.49)	2.68	(0.55)	2.92	(0.94)	3.62	(0.71)	2.49	(0.76)
	+	3.37	(0.51)	4.06	(1.04)	3.03	(0.86)	3.00	(0.72)	2.21	(0.42)
Female	-	3.74	(0.97)	2.71	(0.60)	4.31	(1.35)	3.99	(0.56)	3.49	(0.54)
	+	3.00	(0.45)	4.42	(1.46)	4.06	(0.80)	4.81	(1.02)	4.93	(1.21)

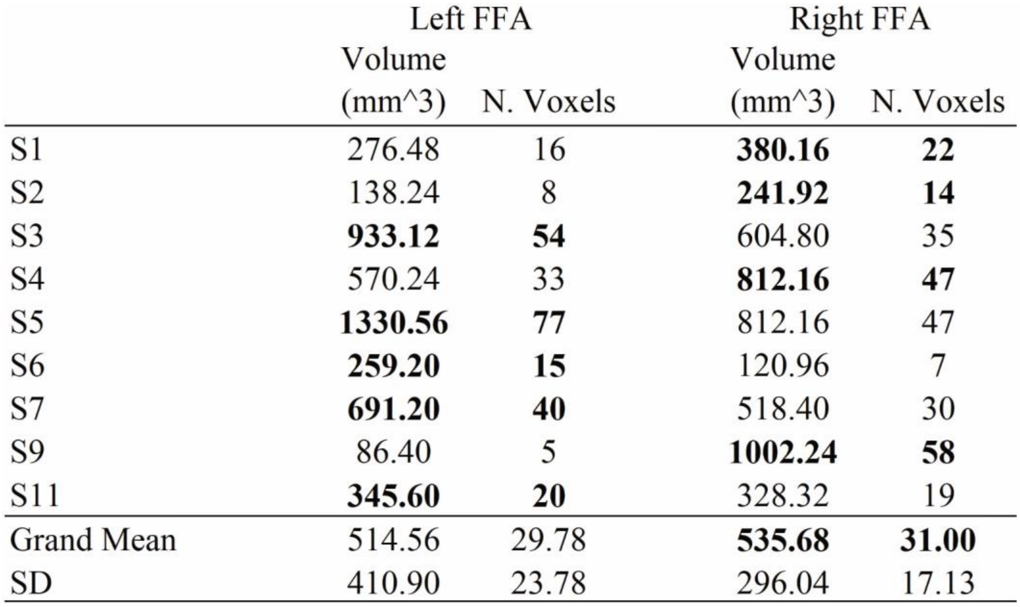
Supplementary Table 3.2. Average naturalness boundaries across the different components for the first 7 participants.

Values shows the means and between-subjects standard deviations of the point at which the first seven participants indicated the face sits on the boundary between natural and physically plausible and unnatural and physically implausible. The units are the number of standard deviations of the loadings of the input set back to into the PCA space from the average face. Note that these are the results from the first 7 participants as these were the results used to generate the stimuli for the MRI experiment.

			S1	S2	S3	S4	S5	S6	S7	S9	S11	**Grand Average**
Component	Gender	Direction	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	**Mean (SD)**
PC1	Male	-	1.83 (0.22)	2.53 (0.20)	1.76 (0.21)	1.55 (0.26)	2.97 (0.28)	2.19 (0.23)	2.13 (0.28)	1.76 (0.26)	2.67 (0.11)	**2.16 (0.48)**
		+	2.65 (0.20)	3.75 (0.31)	3.14 (0.15)	2.83 (0.32)	3.93 (0.79)	3.86 (0.59)	3.47 (0.38)	3.47 (0.51)	4.01 (0.24)	**3.46 (0.49)**
	Female	-	3.66 (0.23)	5.17 (0.63)	3.21 (0.47)	2.69 (0.28)	5.02 (0.68)	3.21 (0.33)	3.20 (0.38)	2.71 (0.42)	4.02 (0.27)	**3.65 (0.92)**
		+	2.92 (0.45)	3.20 (0.49)	2.88 (0.31)	2.10 (0.30)	3.37 (0.37)	3.42 (0.23)	3.11 (0.34)	2.54 (0.24)	3.72 (0.23)	**3.03 (0.49)**
PC2	Male	-	2.57 (0.26)	2.99 (0.37)	2.35 (0.60)	1.76 (0.38)	3.56 (0.29)	2.74 (0.95)	2.77 (0.11)	2.30 (0.47)	3.03 (0.18)	**2.67 (0.51)**
		+	3.60 (0.39)	4.54 (0.59)	3.59 (0.40)	2.55 (0.22)	5.25 (0.69)	5.40 (1.96)	3.45 (0.72)	3.67 (0.74)	5.31 (0.48)	**4.15 (1.01)**
	Female	-	2.69 (0.21)	3.08 (0.24)	2.35 (0.54)	1.80 (0.30)	3.73 (0.17)	2.72 (0.07)	2.60 (0.44)	2.81 (0.32)	2.85 (0.18)	**2.74 (0.52)**
		+	3.66 (0.30)	4.60 (0.19)	3.18 (0.50)	3.22 (0.69)	7.42 (0.93)	4.73 (0.85)	4.16 (0.35)	4.54 (0.37)	4.63 (0.40)	**4.46 (1.26)**
PC3	Male	-	2.35 (0.59)	4.07 (0.52)	2.32 (0.83)	1.75 (0.20)	4.07 (1.15)	2.39 (0.56)	3.49 (0.94)	3.19 (0.78)	3.96 (0.25)	**3.07 (0.89)**
		+	2.46 (0.24)	3.86 (0.35)	2.40 (0.45)	2.43 (1.27)	4.60 (0.21)	2.67 (0.51)	2.78 (0.49)	3.17 (0.33)	4.03 (0.18)	**3.16 (0.82)**
	Female	-	3.59 (0.38)	4.68 (0.76)	3.44 (0.48)	2.87 (0.55)	7.02 (0.99)	3.98 (1.08)	4.62 (0.54)	4.00 (0.34)	5.07 (0.26)	**4.36 (1.21)**
		+	3.61 (0.28)	4.78 (0.72)	3.75 (0.47)	2.88 (0.64)	5.34 (0.71)	4.00 (1.02)	4.07 (0.54)	2.74 (0.43)	4.79 (0.35)	**4.00 (0.87)**
PC4	Male	-	2.99 (0.62)	4.01 (0.26)	3.37 (0.51)	2.95 (0.43)	4.73 (0.43)	4.25 (1.22)	3.03 (0.46)	2.96 (0.73)	4.40 (0.56)	**3.63 (0.71)**
		+	2.32 (0.27)	4.01 (0.78)	2.55 (0.46)	2.22 (0.34)	3.85 (0.67)	3.28 (0.74)	2.76 (0.36)	2.59 (0.65)	3.05 (0.24)	**2.96 (0.64)**
	Female	-	3.91 (0.63)	4.53 (0.53)	3.83 (0.99)	3.18 (0.83)	4.88 (0.35)	3.69 (0.66)	3.91 (0.32)	3.38 (0.26)	4.53 (0.38)	**3.98 (0.56)**
		+	3.71 (0.52)	6.29 (0.58)	4.21 (0.89)	3.83 (1.33)	5.91 (1.12)	5.24 (1.51)	4.50 (0.32)	4.15 (0.71)	5.00 (0.42)	**4.76 (0.91)**
PC5	Male	-	1.79 (0.41)	3.35 (0.58)	2.07 (0.66)	1.41 (0.12)	3.34 (0.30)	2.70 (0.54)	2.79 (0.39)	2.72 (0.61)	2.70 (0.36)	**2.54 (0.66)**
		+	2.24 (0.14)	1.69 (0.57)	2.07 (0.32)	1.92 (0.38)	3.04 (0.46)	2.29 (0.63)	2.19 (0.27)	2.46 (0.44)	2.91 (0.35)	**2.31 (0.44)**
	Female	-	3.09 (0.36)	4.10 (0.47)	2.90 (0.28)	2.92 (0.46)	4.19 (0.79)	3.61 (0.71)	3.64 (0.78)	2.95 (0.63)	3.82 (0.24)	**3.47 (0.52)**
		+	4.36 (0.63)	5.60 (0.81)	3.96 (1.16)	3.61 (1.30)	7.17 (0.82)	5.26 (1.52)	4.54 (0.77)	4.76 (0.82)	6.11 (0.60)	**5.04 (1.12)**

Supplementary Table 3.3. Average naturalness boundaries and within-subject SDs for each of the stimulus types.

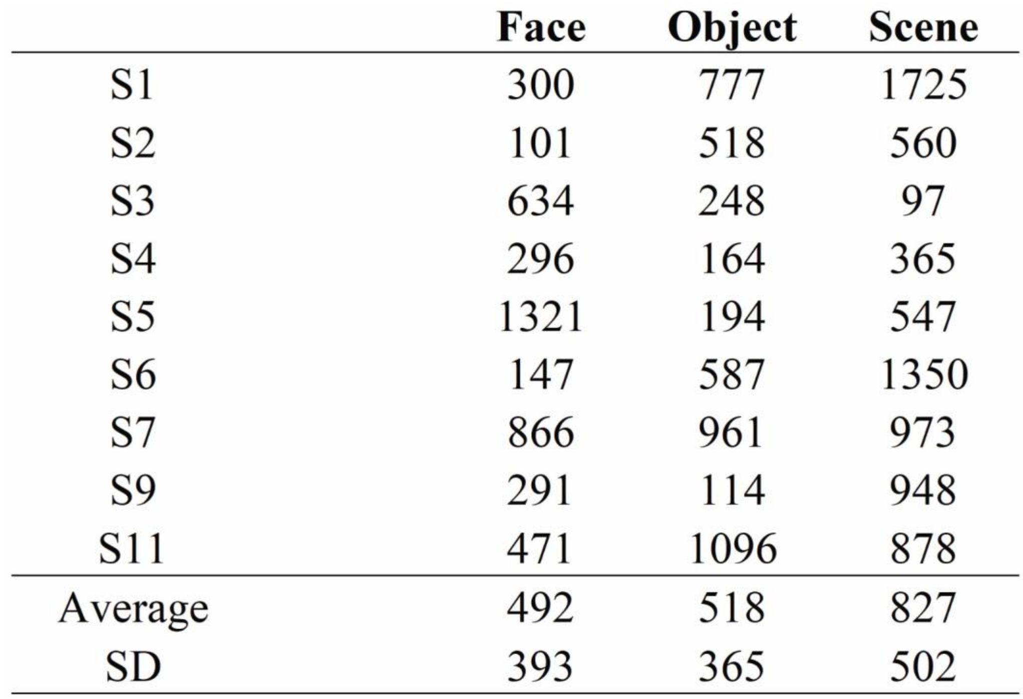
The table shows the average naturalness boundary for each of the 'component' * 'gender' * 'direction' combinations for each participant. SD = the within-subject standard deviation across the 6 repetitions of each stimulus type. The value is the distance from the average face at which point the participants indicated the face is on the boundary of natural variability. Units are in terms of the standard deviations of the loadings of the input image set into the PCA space.



	Left FFA		Right FFA	
	Volume (mm^3)	N. Voxels	Volume (mm^3)	N. Voxels
S1	276.48	16	**380.16**	**22**
S2	138.24	8	**241.92**	**14**
S3	**933.12**	**54**	604.80	35
S4	570.24	33	**812.16**	**47**
S5	**1330.56**	**77**	812.16	47
S6	**259.20**	**15**	120.96	7
S7	**691.20**	**40**	518.40	30
S9	86.40	5	**1002.24**	**58**
S11	**345.60**	**20**	328.32	19
Grand Mean	514.56	29.78	**535.68**	**31.00**
SD	410.90	23.78	296.04	17.13

Supplementary Table 3.4. Volumes (mm^3) and the number of voxels within the left and right FFA.

Bold indicates the hemisphere in which the FFA is larger within each participant. The FFA was defined by presenting voxels that survived FWE correction for faces > objects + scenes on a flat map, and then hand-drawing an ROI around contiguous voxels within the fusiform gyrus.



	Face	Object	Scene
S1	300	777	1725
S2	101	518	560
S3	634	248	97
S4	296	164	365
S5	1321	194	547
S6	147	587	1350
S7	866	961	973
S9	291	114	948
S11	471	1096	878
Average	492	518	827
SD	393	365	502

Supplementary Table 3.5. Number of voxels within face-, object- and scene-selective regions.

The number of voxels defined as face-, object- and scene-selective within the occipitotemporal anatomical ROI defined by the contrasts: faces > objects + scenes; objects > faces + scenes; scenes > faces + objects. Only voxels that survived FWE correction were included. The average shows the mean number of voxels along with the standard deviation (SD).

Excel files of the supplementary tables can also be found in the Supplementary Materials folder.

Chapter 4 A multi-view PCA space for representing and reconstructing dynamic expressions across views

4.1 Preface

The previous chapter assessed one aspect of face space as a linear projection machine by investigating the fMRI response as faces varied in caricature level relative to the origin of an identity-general space. The current chapter addresses a different aspect, addressing how view-invariance for facial motion might be achieved and detailing a bio-inspired model that can represent and reconstruct facial motion across changes in view.

4.2 Introduction

Given the evidence of view-invariant cortical responses (e.g., Chang & Tsao, 2017; Freiwald & Tsao, 2010; Meyers et al., 2015) and the behavioural evidence that adaptation aftereffects transfer across views (Jiang et al., 2006, 2007) it is clear that faces are not processed solely in a view-specific manner (although see Benton et al., 2006; Jeffery et al., 2006). Many strategies for achieving view-invariance exist in computational models, despite still being a stumbling block (Ding & Tao, 2016), but how the brain achieves view-invariance is not yet established. Humans might be able to mentally rotate faces (O'Toole et al., 1998), however it is unknown whether or how the brain performs a 'frontalisation' procedure, either consciously or automatically, or if it holds a 3D or a series of 2D representations. The aim of this chapter is not to formulate a conclusive answer about what method the brain takes. Instead, this chapter expands on a current 2D, view-based model (Beridze, 2021) to assess if a biologically plausible mechanism can be achieved for learning and reconstructing facial motion across viewpoints. Motion provides a large source of within-person variation and cues to identity (e.g., H. Hill & Johnston, 2001; Knight & Johnston, 1997), so uncovering how motion might be integrated across views can enhance our understanding of expression, speech and identity perception and lead to technological advances such as in automatic speech-reading and face recognition systems.

The multiple appearance model created by Beridze (2021) is a PCA-based, simple linear space and projection mechanism. Given that a simple linear projection model (Chang & Tsao, 2017), like PCA, can account for neural responses better than numerous neural networks (Chang et al., 2021) and given inconclusive evidence for a 3D representation, the current work focused on expanding the 2D, view-based methods of Beridze.

In the first model, Beridze (2021) created separate view-specific spaces and tried to map deviation vectors from one space to another, akin to how Cowe (2003) and Nagle et al (2013) mapped expressions across different identities and how Griffin et al (2011) created familial resemblance across genders. This technique is discussed more in the description of our first two models, but to summarise the model only worked well for small viewpoint changes.

The subsequent method by Beridze concatenated multiple viewpoints of simultaneously recorded videos together prior to performing PCA, creating a ‘multiple appearance model’. By projecting one viewpoint onto the components Beridze was able to reconstruct the expression in the other viewpoints, even over large rotations. However, whilst a good starting point, the model was not biologically plausible as it required simultaneous exposure to multiple views. Model 3 of our work attempts to replicate this model and subsequently Models 4 and 5 aimed to improve the biological plausibility of the methods.

As mentioned, retaining expression information and non-rigid motion can also benefit automatic speech-reading systems. With recent advances in technology, wearable lip-reading technologies are becoming closer to reality. As with recognition, these systems can work in various ways, such as learning to lip-read in each view separately (Isobe et al., 2021) or combining all views into one large corpus to learn (Chung & Zisserman, 2017; Isobe et al., 2021). Alternatively, one could learn to lip-read in one viewpoint, reducing computational costs, and transform all inputs to that viewpoint (Lan et al., 2012), akin to the frontalisation procedures discussed in Chapter 1.3.5. Lan and colleagues (2012) found that lip-reading is best at ~30^{o} from frontal for both humans and computer lip-reading systems so devised a view-transformation model to rotate the mouth region, presented in any view, to 30^{o}.

Although automatic lip-reading is not the focus, the current work utilises and expands on some the methods used by Lan and colleagues (2012) and thus may benefit view-transformations for automatic lip-reading. Firstly, Lan and colleagues (2012) cropped to the mouth where our work utilises the whole face. Extraoral regions can enhance the accuracy of automatic visual speech recognition software (Y. Zhang et al., 2020) and vocal tract structure during speech correlates with some extraoral regions (Scholes et al., 2020) suggesting that other facial regions contribute to effective speech-reading. Scholes and colleagues' actors maintained a fairly neutral expression, but emotional expressions convey additional information about speech content (Shirakata & Saitoh, 2020). The current work also differs from that of Lan and colleagues by striving to achieve biological plausibility and avoiding the need to record from distal views simultaneously (e.g., frontal and profile).

Here, five models are outlined, varying in structure and expanding on the methods of Beridze (2021) and Lan and colleagues (2012), culminating with a two-step model of view-invariant motion processing (Model 5). Model 5 was the most biologically plausible, while still accurately reconstructing facial motion across views.

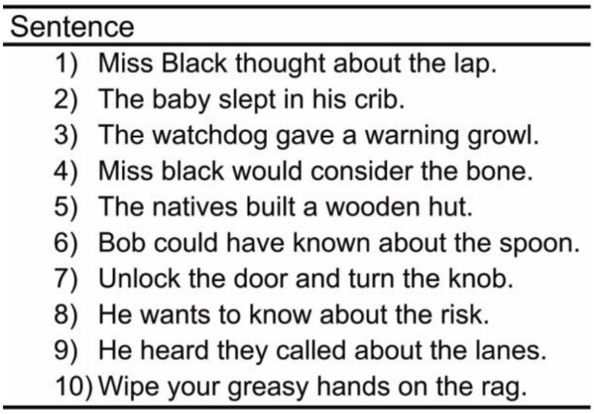
To preface any confusion, the term 'multi-view' is used differently here compared to previous works. In previous models, such as multi-view subspace models, there are separate representations for each viewpoint which are brought into correspondence, for example, by maximising the canonical correlation between loadings on the features across different views (Li et al., 2009; Rupnik & Shawe-Taylor, 2010). The resulting features are considered view-invariant by reflecting corresponding changes in multiple views. In the current work, the term 'multi-view' primarily refers to vectors or components containing multiple viewpoints as separate sections of the vector (e.g., a 20-element vector with a different view every 5 elements).

4.3 Methods

4.3.1 Video capture and preprocessing

4.3.1.1 Video capture

The videos used were acquired for a previous study (Scholes et al., 2020). All 9 actors gave prior consent for their videos to be used. An overview of the video capture is reported here (for more details see Scholes et al., 2020). Videos were simultaneously acquired from 0° (frontal), 22.5°, 45° (3/4 view), 67.5° and 90° (profile) from the left side of the actor's face, from the actor's perspective. Video capture used 5 Grasshopper GRAS-03K2C (FireWire) cameras (PointGrey), placed on a bespoke semicircular camera rig, captured at 30 fps in RGB 24-bit px format, at a resolution of 640 x 480. During video capture, actors repeated 10 sentences 20 times (see Table 4.1), obtained from the Speech Intelligibility in Noise Database (Kalikow et al., 1977).



Sentence
1) Miss Black thought about the lap.
2) The baby slept in his crib.
3) The watchdog gave a warning growl.
4) Miss black would consider the bone.
5) The natives built a wooden hut.
6) Bob could have known about the spoon.
7) Unlock the door and turn the knob.
8) He wants to know about the risk.
9) He heard they called about the lanes.
10) Wipe your greasy hands on the rag.

Table 4.1. Sentences repeated by the actors.
The 10 sentences used by Scholes and colleagues (2020), sourced from the Speech Intelligibility in Noise Database (Kalikow et al., 1977).

4.3.1.2 Cropping and alignment

The videos were first cropped to a square and downscaled to 128 x 128 pixels. The crop was a fixed position approximately centred on the face, allowing natural, rigid motion across the frames. Facial form was normalised to the shape of a reference frame for each view. This used the same procedure as prior studies (Beridze, 2021; Cowe, 2003; Nagle et al., 2013; D. M. Watson & Johnston, 2022). For each actor a reference frame was chosen exhibiting a neutral expression but with the lips parted revealing the teeth. The remainder of the frames were warped to the reference frame (which was then removed) using the Multi-channel Gradient Model (Johnston et al., 1992; 1999). The same frame number was used for each of the simultaneously recorded views.

Warping to the reference frame is comparable to the alignment methods used in automatic face recognition systems (e.g., Craw, 1992; Hassner et al., 2015) and the shape-normalisation used by Burton and colleagues (2016). Simpler alignment methods, such as matching eye and mouth positions are straightforward, but lead to blurring effects where faces and features of different shapes and sizes are superimposed onto one another (Aishwarya & Marcus, 2010), for example the mouth when in motion. More complex procedures, such as using the McGM, or normalising the shape using a greater number of facial landmarks (Burton et al., 2016; Kramer et al., 2017), reduce the superimposition problem and provide shape and motion information that can be utilised.

Manually placing landmarks on each image allows precision but is highly impractical for videos longer than a few seconds. One could use automatic landmarking software such as OpenFace (Baltrusaitis et al., 2018) but, such software works best with frontal images with no occlusion; as noted by Wu and Ji (2019) one of the major challenges to automatically detecting landmarks is large rotations from frontal – a problem we have encountered with OpenFace.

The effect of occlusion can be seen for instance in the Multiview Active Shape Modes outlined by Milborrow et al (2013). The software erroneously assumed that one eye was occluded through a rotation in pose causing it to estimate the face was pointing in the opposite direction and thus the estimates of the landmarks were spatially flipped compared to their actual locations. The eye was occluded by hair.

Landmarking software can also struggle with the intricacies of asymmetric motion. From our own experience, OpenFace assumes the lips move somewhat symmetrically, and as such fails to properly capture asymmetric movement.

In contrast, the McGM is a biologically inspired gradient based motion estimation algorithm that is agnostic to facial landmarks. It also provides a much denser description of shape, with a value for each pixel rather than a value for each landmark. A short summary is outlined by Nagle and colleagues (2013). Because the McGM is not based on landmarking, it does not have some of the same constraints, although there are still issues with occlusion if the occluding object only transiently comes into view, such as a hand itching a nose. The McGM performs the alignment unsupervised, provided there is not too much variation between the frames and the template.

The alignment provides a warp vector field with two, 128 x 128 components for each frame, one for horizontal (X) translations and one for vertical (Y) that move the pixels in the aligned frames back to their original locations. The outputs were therefore each 128 x 128 x 5 (Xwarp + Ywarp + RGB) x *N* where *N* is the number of frames minus the reference frame (see Figure 4.1a).

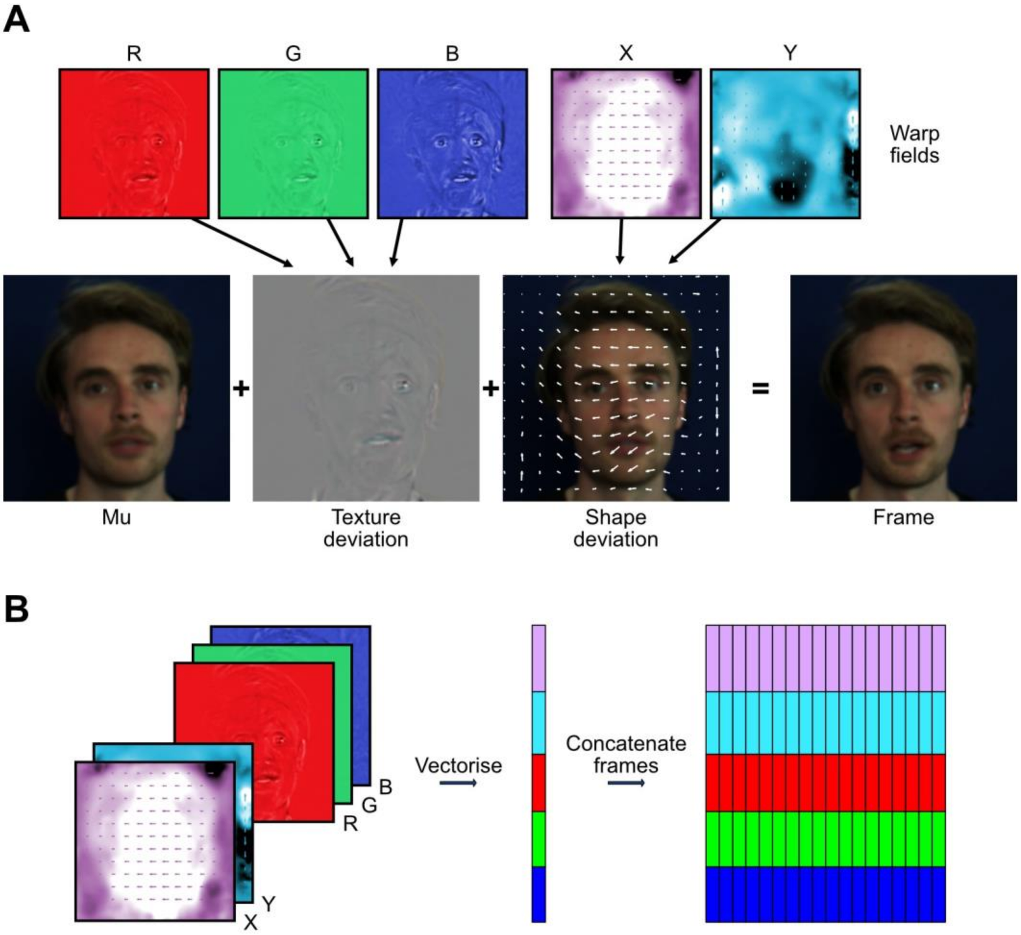


Figure 4.1. Warping and vectorisation.

(**A**) Depiction of warping, showing how to reconstruct a given frame from mu, the RGB texture and XY shape deviations. In X and Y, lighter regions show leftwards and upwards deviations respectively, darker regions show rightwards and downwards deviations. Arrows show the direction and magnitude of the pixel displacement. (**B**) Vectorisation of a single frame and concatenation of multiple frames. In both plots, the colours of the RGB fields have been exaggerated for display purposes only.

The X and Y components are essentially horizontal/vertical deviations embedded in a mesh grid (see the function *meshgrid* in MATLAB). The mesh tells the reconstruction script how to interpolate one image to make another. In a perfect mesh grid, all pixels in a new image are extracted from their original positions, so the image stays the same. To induce a pixel displacement, say in X, values are either increased or decreased to induce a leftward or rightward shift respectively. For a leftward shift, you need to pull the pixels from the right in the original image.

Following warping, frames were vectorised, with each vector containing 81,920 elements (128 x 128 x 5). The first 32,778 contained the shape (warp) values, the remaining 49,152 contained the texture (RGB) values (see Figure 4.1b). Each view's matrix was therefore 81,920 x *N*.

4.3.1.3 Detecting and correcting frame asynchrony

For the models to be accurate the cameras needed to record in perfect synchrony. Unfortunately, they did not, with two recording marginally faster than the others. During recording of ~20,000 frames, two views would end ~4 frames ahead. Whilst a miniscule discrepancy in frame rate, even 2 frames can be the difference between the mouth being closed and nearly fully open.

Frame rate discrepancies and dropped frames first needed identifying in the absence of frame-by-frame timestamps. Over the frames for each view, the vertical position of a landmark (landmark 56) on the left side of the lower lip (from the actor's perspective) was tracked using OpenFace 2 (Baltrusaitis et al., 2018). This was one of the most dynamic landmarks that could be tracked in all views.

To assess for dropped frames or differences in frame rate, the vertical positions in views 2-5 were compared to view 1 (0^{o}) using a sliding window. A sliding window of 100 frames was passed along the videos (step = 1). At each step the correlation in the y-positions across the window was calculated between each view (views 2-5) and 0^{o}, with offsets of -5 to +5 frames introduced. For each step, the offset with the highest correlation coefficient was determined, providing an index of how (a)synchronous the view was from frontal. Synchronous videos should have the highest correlation with no offset throughout.

These offsets were plotted over the duration of the videos. A linear drift was visible, typically for views 2 and 5 whereby the frames would start in, but then drift out of, synchrony with 0^{o}. During this drift there were extended periods where the highest correlation alternated between two offsets, before settling on the second. This suggested asynchrony by a fraction of a frame, leading to uncertainty in the most correlated offset. In contrast, frame drops were shown by a more sudden change. Visual inspection of sudden movements such as blinks across the views confirmed a small discrepancy in the frame rate rather than many dropped frames, other than one dropped frame in one camera for one actor. Sudden movements were assessed at the start of the videos, the end, and around the transition points between offsets.

Frame rate discrepancies were resolved by interpolating the desired number of frames from view-specific PCA spaces. For each view, a separate PCA space was made using all frames. The frames were projected into the space and the loadings onto each component calculated. The loading trajectories were then interpolated using MATLAB's *interp1* function with *spline* interpolation. For example, if view 1 had 20,000 frames, and view 2 ended 4 frames ahead, 20,000 frames were interpolated from the loading trajectory between frames 1 and 19,996 in view 2's space. This allowed fractional discrepancies to be corrected. Here the PCA spaces only retained ≥100 components and ≥90% of the variance, so some data loss is possible through this method, but this can also, beneficially, remove some noise prior to making the multi-view space.

Some frames were also removed due to excessive head movements that caused artefacts in the warping, such as dropping the head substantially, or where the face was obscured by the hand. These were omitted when making the separate view spaces, included when calculating loading trajectories to avoid discontinuities in the interpolation, and were then removed afterwards. Once the frames had been synchronised, the PCA models were created.

4.3.2 Projection and scaling

Often, principal components are unit vectors, however this was not always the case here. In some such cases, we used the full equation for projection, but in others we ignored the denominator. In other words, we took the inner product between the deviation vector D and component EF and did not divide that by the squared length of EF.

In some models eigenframes contained 5 viewpoints and the aim was to reconstruct all views from a single view. We therefore projected only one viewpoint into the space. One strategy for this was to retain the whole eigenframe and replace undesired parts of the input vector with zeros as in equation (4.1), where e is the number of elements in the original multi-view vector, D, and pD_v is the partial vector for the view v where the rest of the vector is replaced with zeros. Here, view 2 of 5 is isolated.

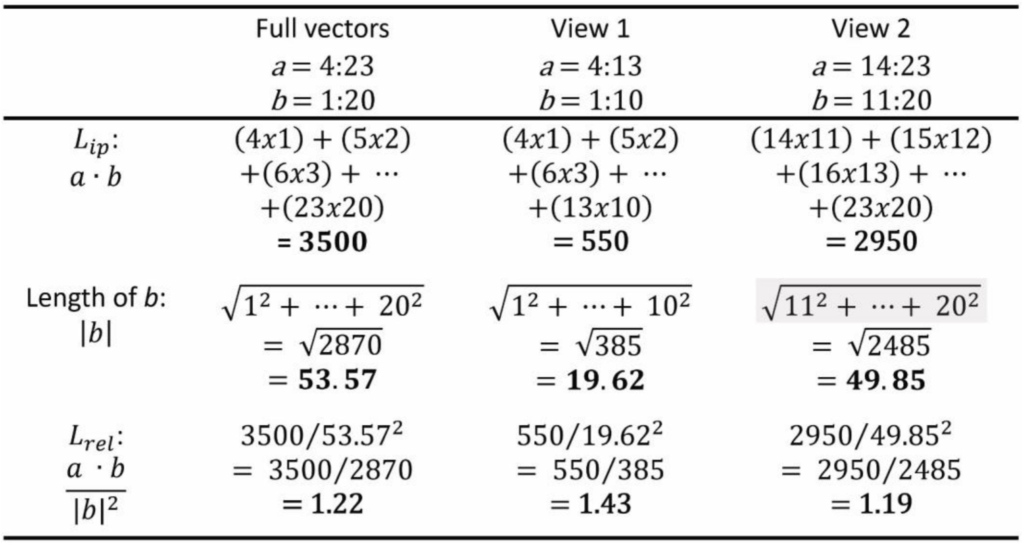
$$\boldsymbol{pD}_v = [\emptyset_{\mathbf{1}} \cdots \emptyset_{\frac{e}{5}}, \boldsymbol{D}_{(\frac{e}{5})+1} \cdots \boldsymbol{D}_{2(\frac{e}{5})}, \emptyset_{2(\frac{e}{5})+1} \cdots \emptyset_{e}] \tag{4.1}$$

This strategy has been used previously (Beridze, 2021; Scholes et al., 2020), but there are two reasons why truncating D (tD_v) and EF (tEF_v) to the input view may be a better strategy. The first is minimising computational demands. It is less demanding to project the 81,920 elements for a single view (tD_v) onto 81,920 elements of the truncated components (tEF_v), than all 409,600 elements in pD_v onto all elements of EF. As the full components are unit vectors the inner product of $tD_v \cdot tEF_v$ is equivalent to $\frac{pD_v \cdot EF}{|EF|^2}$.

The second reason regards scaling. Both Beridze (2021) and Scholes et al (2020) found that the inner product loadings from partial projection (padding the input with zeros) were much lower than when projecting in the full vectors. This is due to using fewer elements to calculate the inner product. An alternative option would be to truncate the vectors and eigenframes and perform the full projection calculation. tEF_v is no longer a unit vector, so the projection can then be calculated as the magnitude of $tD_v \cdot tEF_v$ relative to the magnitude of $|tEF_v|^2$. As outlined shortly, this method is more stable across changes in vector magnitude and so might provide a method for scaling the loadings.

As we refer to these methods of projection frequently, it is first worth outlining some terms to summarise these methods. When only the inner product between D and EF (or tD_v and tEF_v) is calculated, this is referred to as the *inner product loading* (L_{ip}). It is *not* scaled relative to the length of the vector being projected onto. When the denominator is included and $tD_v \cdot tEF_v$ is divided by $|tEF_v|^2$, this is referred to as the *relative loading* (L_{rel}), as it is the magnitude of the projection *relative* to the magnitude of the vector being projected onto. When the vector being projected onto is of unit length, $L_{rel} = L_{ip}$.

The relative loading will be the same regardless of the magnitude of the vectors, so long as the angle between the vectors and the ratio of the magnitudes are kept constant. In contrast, the inner product loading will vary. As an example, consider two vectors, a (a new frame) which is 4:23 and b (a component) which is 1:20. For this example, both are 2-view vectors. View 1's loadings can be calculated by projecting the first 10 elements of a onto the first 10 elements of b. The same can then be calculated for view 2 with the last 10 elements. The inner product and relative loadings are outlined in Table 4.2. The relative loading is the relative magnitude of the projection of a onto b and is therefore much more stable to variation in the magnitudes of the vectors. As a result, the relative loading is explored as a method for scaling the loadings in the multi-view PCA models.



	Full vectors $a = 4{:}23$ $b = 1{:}20$	View 1 $a = 4{:}13$ $b = 1{:}10$	View 2 $a = 14{:}23$ $b = 11{:}20$
L_{ip}: $a \cdot b$	$(4x1) + (5x2)$ $+(6x3) + \cdots$ $+(23x20)$ $= \mathbf{3500}$	$(4x1) + (5x2)$ $+(6x3) + \cdots$ $+(13x10)$ $= \mathbf{550}$	$(14x11) + (15x12)$ $+(16x13) + \cdots$ $+(23x20)$ $= \mathbf{2950}$
Length of b: $\lvert b \rvert$	$\sqrt{1^2 + \cdots + 20^2}$ $= \sqrt{2870}$ $= \mathbf{53.57}$	$\sqrt{1^2 + \cdots + 10^2}$ $= \sqrt{385}$ $= \mathbf{19.62}$	$\sqrt{11^2 + \cdots + 20^2}$ $= \sqrt{2485}$ $= \mathbf{49.85}$
L_{rel}: $\frac{a \cdot b}{\lvert b \rvert^2}$	$3500/53.57^2$ $= 3500/2870$ $= \mathbf{1.22}$	$550/19.62^2$ $= 550/385$ $= \mathbf{1.43}$	$2950/49.85^2$ $= 2950/2485$ $= \mathbf{1.19}$

Table 4.2. Example comparison of inner product and relative loadings.

Example showing how the relative loading of a projection is less susceptible to differences in magnitude within a vector compared to the inner product loading.

4.3.3 Measuring reconstruction accuracy

Multiple ways to quantify reconstruction accuracy exist, each with their own advantages and disadvantages. Broadly, quantification can be split into the *feature* being measured and the *measure* used. The *feature* options are image-level, deviation-vector-level, and single space loadings. The *measure* options considered here are the sum of squared error (SSE) or Euclidean distance (ED, = √SSE), and the correlation coefficient (Pearson's *r* or Fisher's Z transform). Generally, for comparing models, reconstruction accuracy was averaged across frames, providing one value per input and reconstructed view per actor. These were often averaged across view providing one value per model for frames used in training and for frames in the test set, with a separate value for single-view and multi-view inputs (whether a single view was projected into the space or all views simultaneously). The average of the single-view inputs was then compared across models using ANOVAs and t-tests.

The conversion from Pearson's *r* to Fisher's Z allowed parametric comparisons by transforming the data to fit a normal distribution, however in doing so correlation coefficients of ±1, or close enough to 1 with rounding errors, resulted in infinite values. If ignored, then reconstruction accuracy would appear worse than it was. To mitigate this, infinite z-values were replaced with ±9.557, corresponding to a Pearson's *r* of 1-1x10^{-8}.

4.3.3.1 Image-level statistics

For image-level (or pixelwise) statistics, each reconstructed frame was vectorised, and compared to a vectorised version of the actual frame by calculating the correlation or ED between them. This was performed for each frame, and each input and reconstructed view, separately.

There are some limitations of using correlation here. Firstly, small spatial translations of the reconstructions, e.g., the whole head being shifted slightly leftwards, can result in a decrease in the reconstruction accuracy despite being otherwise accurate. Secondly, as discussed by Scholes et al (2020), invariant background pixels inflate the correlation coefficient. Thirdly, correlation is not sensitive to global, uniform changes in magnitude, but this can be beneficial as uniform changes on all pixel intensities have no specific effect on the face. Nevertheless, a direct comparison between models based on the correlation is still informative.

The ED, in contrast, is sensitive to changes in pixel magnitude, and so provides another measure of reconstruction accuracy. It is less sensitive to the background pixels, and can be compared across models, but the result is relatively arbitrary. What value constitutes a ‘good’ reconstruction? Zero is perfect, but what is the upper limit for acceptable reconstructions? Nevertheless, it too is informative for comparing models.

4.3.3.2 Deviation-vector similarity

To circumnavigate some of the issues of using pixelwise comparisons, deviation vectors can be compared using the same measures as above. Because the textures should be spatially aligned in the deviation vectors, the correlation between the texture portions of the vectors should be less affected by spatial translations, even if present in the reconstructed image.

If the spatial translation is caused by an over-estimation of the warping, then this might also not contribute much to the correlation coefficient if there is a global change in the magnitude rather than the pattern. As with the pixelwise comparisons however, the background pixels in the deviation can also artificially inflate the correlation coefficient. In contrast, ED would again be less sensitive to the background pixels, but would be sensitive to any over or underestimations of the pixel displacements and texture values.

4.3.3.3 Single view/separate space loadings

An alternative approach, similar to the one used by Scholes et al (2020) is to compare actual and reconstructed frames within a PCA space. Scholes and colleagues projected one modality into their combined video and MR slice space and reconstructed the other. They then projected the full, joint-modality reconstruction back into the space, alongside the veridical joint-modality vector, and compared the loadings between the projections. Here, the equivalent would be comparing reconstructed and actual multi-view vectors in a multi-view space, however, not all the models have a multi-view space. Instead, reconstructions and actual frames were split by view, and projected into separate spaces for each view.

These loadings can again be compared with correlation and ED. A benefit of using the separate space loadings is that the result is no longer inflated by the background pixels. Note, however, that the first layer of Model 5 is comprised of these separate spaces. Therefore, the loadings of the actual frames and the reconstructed frames of the same view are identical, so same-view reconstructions are ignored for statistical comparison. This problem does not apply to cross-view reconstructions, and does not apply when looking at the second, multi-view layer.

4.4 Model 1 - "Daisy chain" model

4.4.1 Model aim and overview

This first model aimed to test whether mapping deviation vectors across neighbouring views' spaces would provide an alternative method over directly projecting across large rotations in yaw (Beridze, 2021). As with previous view-based methods (Pentland et al., 1994), Beridze first constructed a separate space for each viewpoint. Using the same procedure as previous work (Cowe, 2003; Griffin et al., 2011; Nagle et al., 2013), Beridze projected deviation vectors in image space from one view directly into another's space, reconstructing similar views well but struggling more with larger rotations. This model instead only projected deviation vectors into neighbouring views' spaces. It subsequently reconstructed the deviation vectors in those views before projected them into the next space, and so forth.

One might consider the possibility of projecting a frame from one view into its own space and directly transferring the loadings to other spaces. Unfortunately, it is not that simple. Because the different viewpoints have access to different information about changes in texture and motion, the components can vary across different spaces both in their ordering and in their content. For instance, PC1 for a profile space might reflect the head leaning forwards and backwards, whereas in a frontal space it might reflect the head moving side to side. The components might also combine different movements and textures across the different viewpoints, so no single component in one space might perfectly reflect the action of a component in another. As a result, one cannot simply map the loading of PC1 for frontal onto PC1 for profile as it may represent a different action. One would therefore need to work out the correspondence between components across the views, an idea returned to in Model 5. Previous methods (Cowe, 2003; Griffin et al., 2011; Nagle et al., 2013) circumnavigated this problem by working within image space, meaning the correspondence between the spaces did not need calculating and the axes of the spaces did not need to be oriented in the same direction. This worked because the faces were all frontal and were spatially aligned with a simple affine transformation. Raised eyebrows in one space for example spatially matched raised eyebrows in another, loading onto corresponding components.

The problem with mapping deviation vectors across larger changes in view is the spatial misalignment of features. As deviation vectors are based in image space, the spatial position of features, such as the mouth, varies within the vectors for different views, and this difference becomes more prominent for larger rotations. Hence, reconstructions across larger rotations were worse despite Beridze's attempts to align the views by centrally positioning the features. In the current work the views were coarsely spatially aligned, with internal features approximately central along the x-axis. We did not perform more rigorous alignment partly to highlight the spatial sensitivity of this method. Also, there is the theoretical problem of which features to align; even in the best case it is not possible to perfectly align both eyes and the mouth at the same time. Rigid head motion also separates features across views; a forward movement will not change horizontal eye position at 0° but will at 22.5°.

In the current model, instead of directly projecting across large changes in view, a cascading “daisy chain” approach along the views was developed. For instance, frontal (0°) frames were projected into the neighbouring view’s space (22.5°) where the spatial overlap should be the most consistent, and the frames (and deviation vectors) were reconstructed from this space. They were reconstructed by calculating the loadings of the frontal deviation vectors onto the components at 22.5°, scaling the components by the loadings and then adding the average representation for the 22.5° view. The reconstructed deviation vectors were then projected into the next view (45°) and so forth until all views were reconstructed.

To preface the results, we showed how susceptible the model is to changes in spatial arrangement when deviation vectors are directly projected from one space to another. If the alignment issue can be resolved, then this has the potential to be a biologically plausible model.

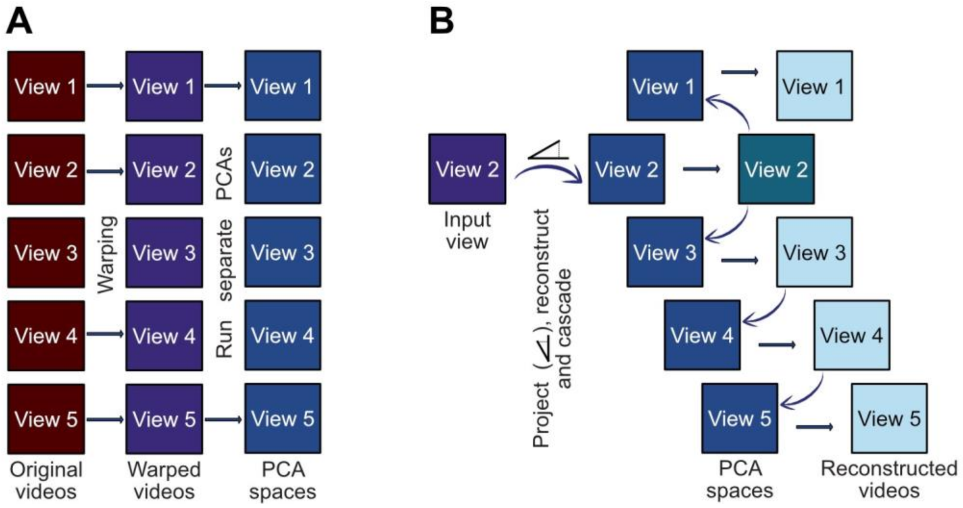


Figure 4.2. Depiction of Model 1.

Depiction of the 'Daisy chain' model (**A**) and how to reconstruct frames across views (**B**). (**A**) Videos (red boxes) are first warped using the McGM to a template frame, each X/Y + RGB image is then vectorised. Purple boxes represent the matrices of warped frames, each with n columns, each represented a frame, and 81,920 rows, depicting the X/Y warp fields and the RGB texture maps. (**B**) Reconstructions across views are made by first reconstructing in the input view and then projecting the reconstructed deviation vectors into the neighbouring view and cascading the process along.

4.4.2 Model construction

An outline of the model is presented in Figure 4.2a. The model was piloted with one actor. As in the previous works (e.g., Beridze, 2021; Pentland et al., 1994), a separate PCA space was made for each view. To create the PCA spaces the first half (n) of the frames were used for training, with the remainder stored for testing. For each view, the mean vector ($\bar{X}_v$ where v denotes the view) was subtracted from the frames creating a matrix of deviation vectors (D_v), each containing 81,920 x n elements. For each view, PCA was performed on D_v using

SVD, returning a matrix of P components where P ≥100, explaining ≥90% of the variance in D_v.

4.4.3 Reconstructing frames across views

An outline of how to reconstruct frames across views is provided in Figure 4.2b. Reconstructions were first made in the input view's space. The frames were projected into the space and loadings on the components calculated. Deviation vectors were then reconstructed as weighted sums of the components using the loadings.

The reconstructed deviation vectors were then projected into the neighbouring spaces. For example, input view 3's reconstructed deviation vectors were projected into, and reconstructed from, the spaces for views 2 and 4. Those reconstructions were then projected into the spaces for views 1 and 5 respectively.

Once the daisy chain process was complete, the average vector ($\bar{X}_v$) for each view was added to the deviation vectors and the result transformed back into 128 x 128 x 3 (RGB) images using the X and Y warp fields to de-warp the frames. The warping removes (but stores) the rigid and non-rigid head movement, and therefore de-warping adds the motion back in.

4.4.4 Results and discussion

Reconstructions made using the daisy chain model can be seen in Figure 4.3 and in the supplementary videos. Same-view reconstructions are good, yet cross-view reconstructions are more variable due to spatial misalignment when moving across neighbouring views. In the example in Figure 4.3, reconstructing frontal or 22.5^{o} frames from each other works well, as does 67.5^{o} from 45^{o}, as the features overlap more, yet reconstructing 45^{o} from 22.5^{o} does not due to a slight spatial misalignment of the features. For instance, the mouth at 45^{o} is slightly further toward the left of the frame than at 22.5^{o}.

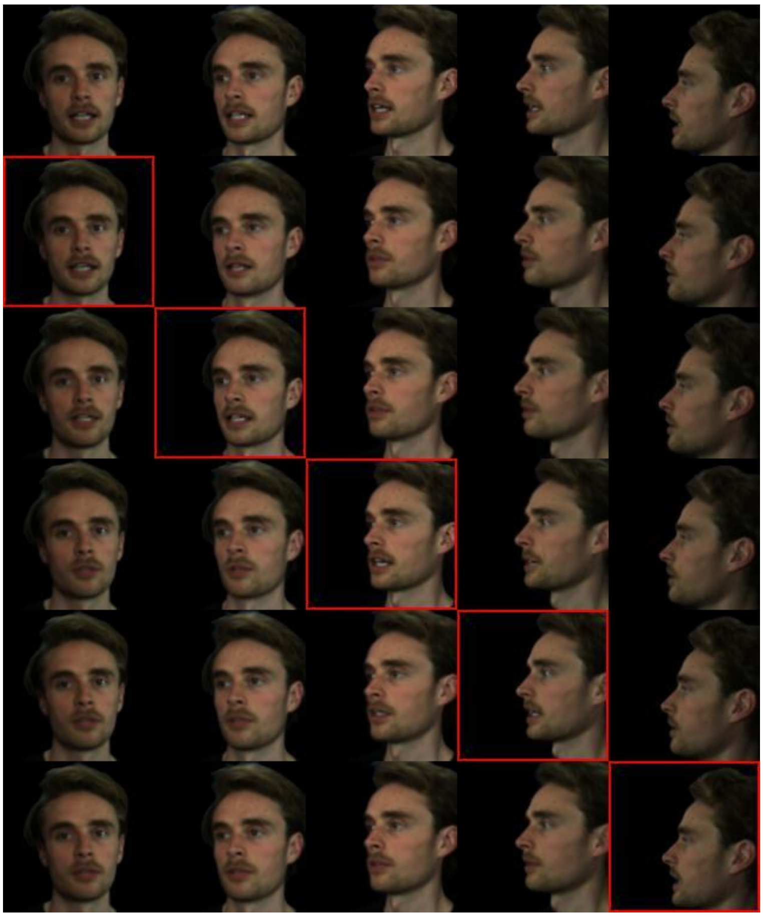


Figure 4.3. Example reconstructions using the daisy chain model. The top row shows the actual frames. The subsequent rows show the reconstructions made using the view bounded by the red box as the input view.

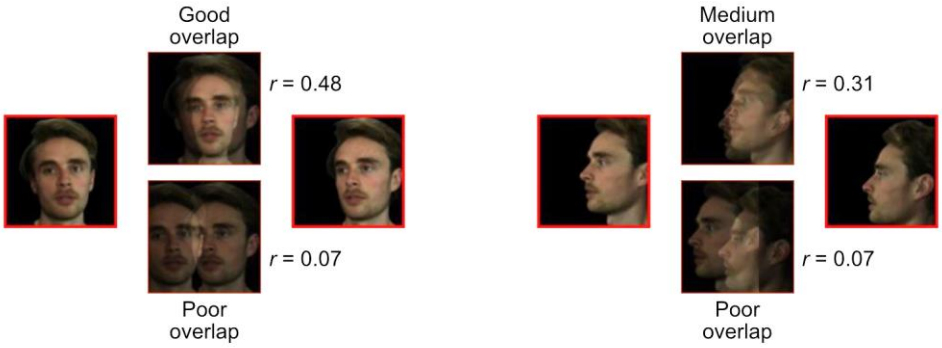


Figure 4.4. The effect of spatial position on reconstruction accuracy.
Demonstration of how reconstruction accuracy varies with the spatial overlap between the input vectors and the PCA components. In each set the central images show the spatial overlap between the left and right images when no spatial offset is introduced (top) and when the left view is horizontally offset by 30 pixels (bottom). The *r*-values show the average loading similarity between projecting the deviation vectors from the left view into the right view's space, compared to the projecting in the right view without any spatial manipulation. For example, in the left set, it is the loading similarity between projecting the reconstructed deviation vectors from view 1 into view 2's space, compared to projecting in view 2's deviation vectors.

To assess this problem further, reconstruction accuracy was assessed when a spatial shuffle was introduced (see supplementary videos), mimicking the effect of cropping the frames to different regions during preprocessing. In between the input and neighbouring views, the horizontal and vertical position of the face was manipulated by moving the input view in increments of 10 pixels horizontally and vertically. The loading similarity (Pearson's *r*) was compared between the cross-view reconstructions and same-view reconstructions. For example, the loading similarity in view 2's spaces when a) view 2 was the input view and b) when view 1 was the input view and the deviation vectors were projected into view 2’s space. As shown in Figure 4.4, the correlation coefficient is heavily dependent on the spatial overlap.

This problem of spatial position can be seen further by exaggerating the loadings after projecting across views. An example is when the mouth in one view overlaps with the cheek in another. The PCA spaces can only reconstruct information present during training, so cannot reconstruct teeth in the wrong place. However, the inner product is higher for components where the area of the cheek matching the position of the teeth is lighter, such as when less in shadow, compared to components where the cheek is darker.

We placed the internal features roughly within the centre of the image but did not align the features further across views unlike Beridze (2021). Such alignment would improve the reconstruction accuracy as demonstrated when a spatial shuffle was introduced. However, as already described, this will not give the perfect amount of overlap in the best case and motion can also separate features across views.

An element also not yet included in this model which we aim to incorporate, is view-invariant components. These components must contain information about all 5 views to behave in a view-invariant manner and stand a chance of mimicking view-invariant neurons (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Meyers et al., 2015).

In summary, this model shows some promise, but is ultimately too sensitive to the spatial overlap between the deviation vectors and the PCA components. It also does not sufficiently recapitulate view-invariance seen in the brain.

4.5 Model 2 – Single slot model

4.5.1 Model aim and overview

The aim of the previous model was to build separate, view-specific spaces and perform cross-view reconstructions by projecting deviation vectors across neighbouring spaces. One limitation was the absence of any view-invariant units. The aim of Model 2 was to explore the possibility of a single, view-invariant space with overlapping representations for the different views. Figure 4.5 shows a summary of the model.

Griffin and colleagues (2011) made separate spaces for male and female faces, but observed adaptation aftereffects that transferred across genders, suggesting a shared neural population. This raises the possibility of genders being represented as clusters within one space, allowing for local gender-specific means as well as the transfer of aftereffects across genders. The transfer of adaptation aftereffects for familiar faces across views also suggests a shared neural population (Jiang et al., 2007). Therefore, rather than creating separate spaces for each view, requiring separate neural populations, all views were entered into one PCA. Separate views were then treated as separate clusters around view-specific means within the space.

As will be discussed, later models created a single space but separated views into different ‘slots’, with each input vector consisting of 5 unique portions, one for each view. Instead, this ‘single slot’ model treated different views as different timepoints. The model then learned how the separate views were clustered in the space, and identified local, view-specific means, allowing for both norm-based representations within views and cross-view adaptation aftereffects (Jiang et al., 2007, 2009).

This model is loosely similar to the Multiview subspace models reviewed by Ding and Tao (2016) in that we hoped to house the separate views within a common space. The difference is that the models discussed by Ding and Tao had separate spaces for each view which were, for instance, rotated using canonical correlation analysis so that the dimensions overlapped sufficiently to count as a common space. Instead, we were trying to find and use local clusters for each view *within* a pre-existing common space.

We were sceptical about this model, both expecting there to be and indeed finding superimposition of features in incorrect places, such as teeth superimposed onto the cheek when a profile view is reconstructed from frontal. But, formally proving that methods *do not* work can be as helpful as proving that others *do*.

4.5.2 Model construction

As with Model 1, the PCA space was created using the first half (n) of the frames for one actor. The views were concatenated together such that each view occupied a different portion of 'time,' as if seeing one view and then the next, creating one 81,920 x $5n$ matrix (X). Although, the temporal sequence is immaterial for PCA, each vector is just a sample. For mean centring the matrix, averaging across all frames in X would result in a morph of all 5 views ($\bar{X}$), and super-subtraction of features in the wrong place in different views when $\bar{X}$ is subtracted. To avoid this, view-specific averages ($\bar{X}_v$) were subtracted prior to concatenating the views. Zero-centring each view meant the concatenated matrix was also zero-centred. PCA was then performed on this matrix.

4.5.3 Identifying view-centres

To learn where each view sits in the space, the loadings (L) of the training frames onto the components/eigenframes for each view were calculated. The average loading on each eigenframe was calculated for each view separately to determine the positions of the view-centres, or local means (see Figure 4.5b)

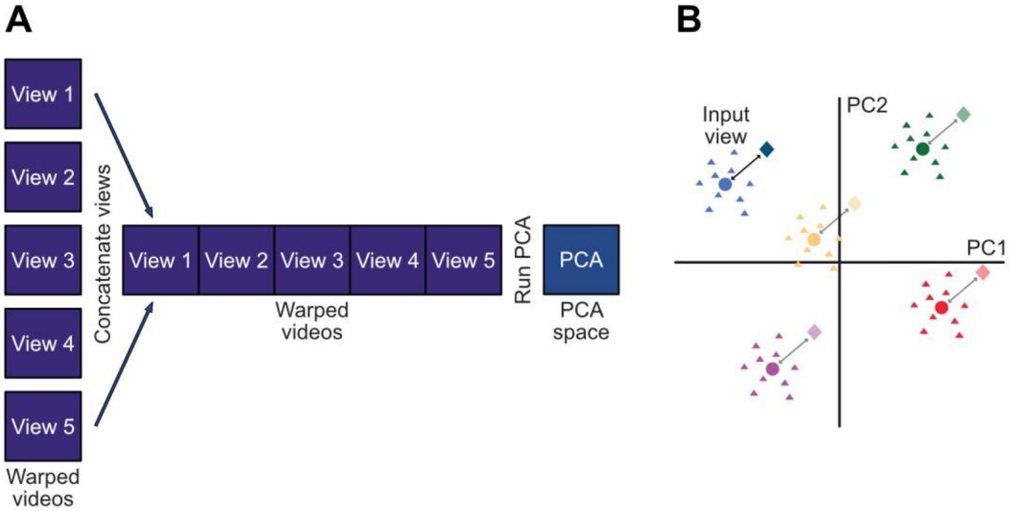


Figure 4.5. Depiction of Model 2

(**A**) An outline of the single slot model. Each box depicts a matrix. The purple boxes depict the warped frames. Within each purple box there are n frames, each represented by a column containing 81,920 rows, containing the X/Y warp fields and the RGB texture maps. The matrices for the different views were concatenated in time, giving an 81,920 x $5n$ matrix. PCA was then performed on this matrix. (**B**) An illustration of the method for calculating view-centres and reconstructing across views. Each colour represents a different view. Triangles represent the loadings of individual training frames on PCs 1 and 2. Circles represent the view-centres, determined by averaging the loadings of the triangles. The blue diamond represents the loadings of a test frame. The relative difference from the view-centre is calculated (the arrow) and added to the other view-centres (other diamonds) to reconstruct across views. This does not depict actual data.

4.5.4 Reconstructing frames across views

For reconstructing frames across views, we used methods conceptually similar to those of Griffin et al (2011), in that we could describe different views relative to one larger space, and therefore map a difference vector from one local mean to another. However, as discussed with the previous model, deviation vectors calculated in *image space* are not compatible across views. Therefore, *difference* vectors for cross-view reconstructions were calculated in terms of the relative difference between the loadings in the space for individual frames and the view-centres.

To reconstruct a frame from one view in another view, the mean vector ($\bar{X}_v$) for the input view was subtracted from the frame giving a deviation vector (D_v), which was projected into the space. This provided loadings relative to the origin (L). To determine the position relative to this vector's view-centre, the loadings of the view-centre (L_{vc}) were subtracted from L, giving a difference vector (d). To reconstruct across views, d was added to the other view-centres (see Figure 4.5b). For each view, the eigenframes were scaled by these new loadings and summed giving a deviation vector for each view, to which the mean vector ($\bar{X}_v$) was added. This vector was then de-warped and reshaped back into a 128 x 128 x 3 (RGB) image.

4.5.5 Results and discussion

The aim of this model was to create a single space with the hope to create view-invariant units. This model captured some changes across views (see Figure 4.6) however, it was possible to see unnatural distortions and the problem of cross-view superimposition was still evident, so this model is also not suitable. For example, see the superimposition of the mouth on the cheek in Figure 4.6 and in the supplementary videos.

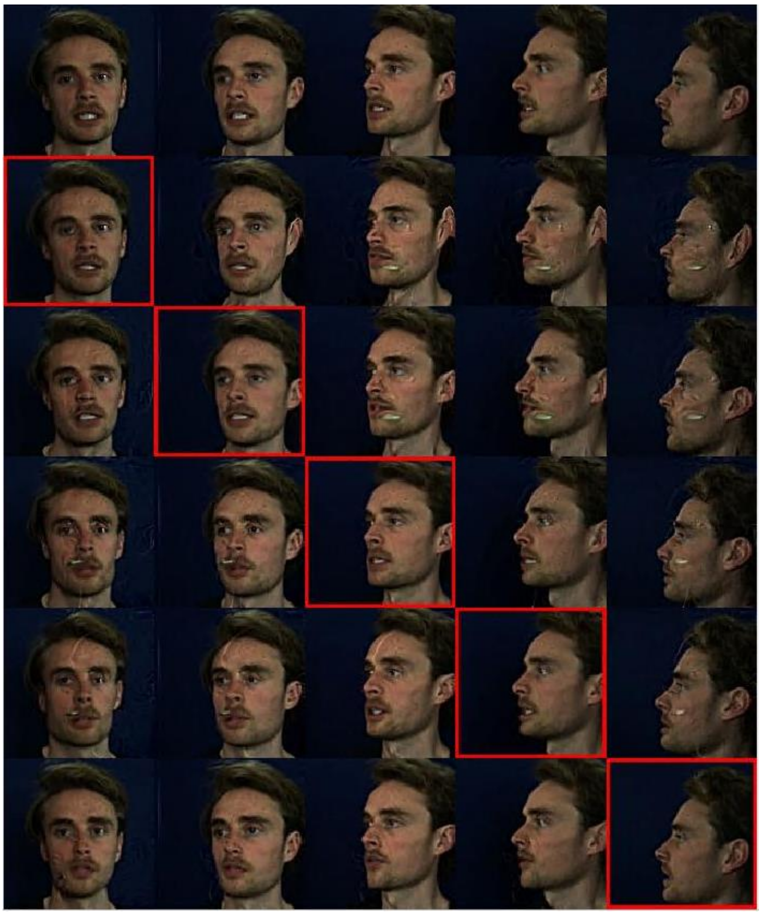


Figure 4.6. Reconstructions from Model 2

The top row shows the original frame for each view. The subsequent rows show the reconstructions for each view made from projecting the view bounded by the red box into the PCA space. As can be seen, these reconstructions suffer from the superimposition of different images, such as the teeth appearing on the cheeks.

The components did not capture correlated movements across views well, so other than large global movements, they tended to be quite view-specific. They also contained unnatural distortions and superimposed features. This can be seen in Figure 4.7 and in the supplementary videos. Note how in PC5 the actions differ across the views and the teeth are superimposed onto the cheek in more profile views.

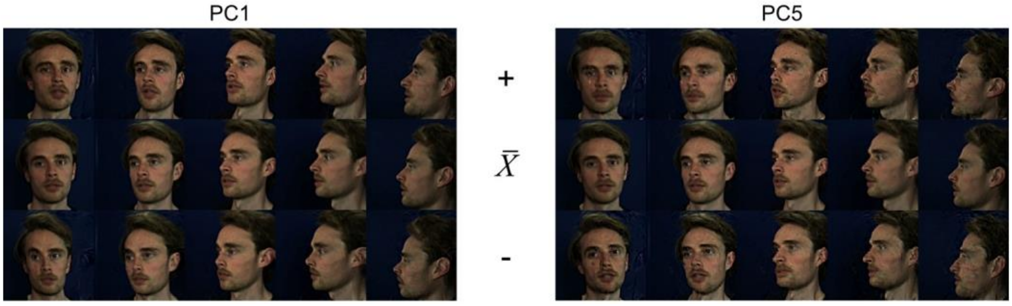


Figure 4.7. Example components of single-slot model.
Depictions of principal components 1 (left) and 5 (right) of the single-slot space for one actor. In each set the middle row represents the local mean for each view. The top and bottom rows represent scaling the components by +3SD and -3SD respectively, where SD is the standard deviation of the loadings of the training set on the components.

4.6 Model 3 – Multi-view, simultaneous entry

4.6.1 Model aim and overview

In Model 2, all views were entered within a single ‘slot’. Model 3 instead tried to recreate the multi-slot approach from Beridze’s (2021) multiple appearance model, where each view has its own slot (own set of pixels, see Figure 4.8). As with the previous models, it assumes view-detection has already occurred and thus inputs could be fed into the correct slot. The same approach was taken by Scholes et al (2020) who concatenated videos with MR slices of the vocal tract such that each ‘timepoint’ contained two modalities.

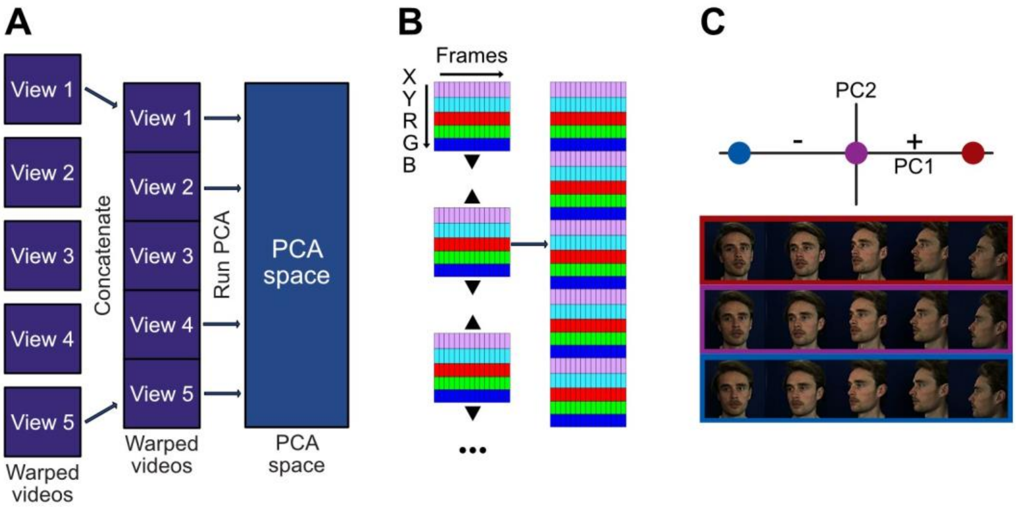


Figure 4.8. Depiction of Model 3

(**A**) Outline of the multi-view simultaneous entry model. Vectorised versions of the frames were concatenated such that each frame contained information about all 5 views, before performing PCA on the multi-view matrix. (**B**) Concatenation of 15 vectorised frames across the 5 views, showing the X and Y warp fields (lilac, cyan) and RGB textures. Each individual box represents a column vector with 16,384 rows (128 *x* 128). (**C**) Example component from PCA. Purple (middle) represents the origin. Red (top) and blue (bottom) show the reconstructed multi-view images from a positive and equidistant negative position along PC1 respectively.

4.6.2 Model construction

See Figure 4.8a for a summary of the model. Whereas in Model 2, the different views were concatenated such that they occupied different periods of 'time', here they were concatenated together such that each 'timepoint' contained all five views, creating one 409,600 x n matrix (matrix X, see Figure 4.8b). Because the videos were captured simultaneously, each view expressed the same action within a given frame. Figure 4.8c shows an example of PC1 for one actor. This model was created and tested for each of the 9 actors.

The first half (n) of the frames were used for training. The mean multi-view vector $(\bar{X})$ was subtracted from the concatenated frames giving a zero-centred deviation matrix D. This is outlined in equation (4.2) where the first subscript element is the view and the second the frame, where n is the number of frames. For simplicity, x, y, r, g, and b each represent a column vector containing all 16,384 (128 x 128) image elements for that channel. As the inputs were multi-view vectors, the eigenframes returned by PCA contained information about all five views.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & & x_{1,n} \\ y_{1,1} & y_{1,2} & & y_{1,n} \\ r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ g_{1,1} & g_{1,2} & & g_{1,n} \\ b_{1,1} & b_{1,2} & & b_{1,n} \\ \vdots & & \ddots & \vdots \\ x_{5,1} & x_{5,2} & & x_{5,n} \\ y_{5,1} & y_{5,2} & & y_{5,n} \\ r_{5,1} & r_{5,2} & \cdots & r_{5,n} \\ g_{5,1} & g_{5,2} & & g_{5,n} \\ b_{5,1} & b_{5,2} & & b_{5,n} \end{bmatrix} \quad \bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{y}_1 \\ \bar{r}_1 \\ \bar{g}_1 \\ \bar{b}_1 \\ \vdots \\ \bar{x}_5 \\ \bar{y}_5 \\ \bar{r}_5 \\ \bar{g}_5 \\ \bar{b}_5 \end{bmatrix} \quad D = X - \bar{X} \tag{4.2}$$

$$e.g., where\ \bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{1,j}$$

4.6.3 Reconstructing frames across views

As each view had its own slot it was not necessary to identify view-centres. To reconstruct frames the mean vector $(\bar{X})$ was subtracted to transform the multi-view images into multi-view deviation vectors (D). For reasons that will become clearer later, the reconstruction process began by projecting the full multi-view vectors into the PCA space. This provided loadings (L) of each full multi-view vector onto each component.

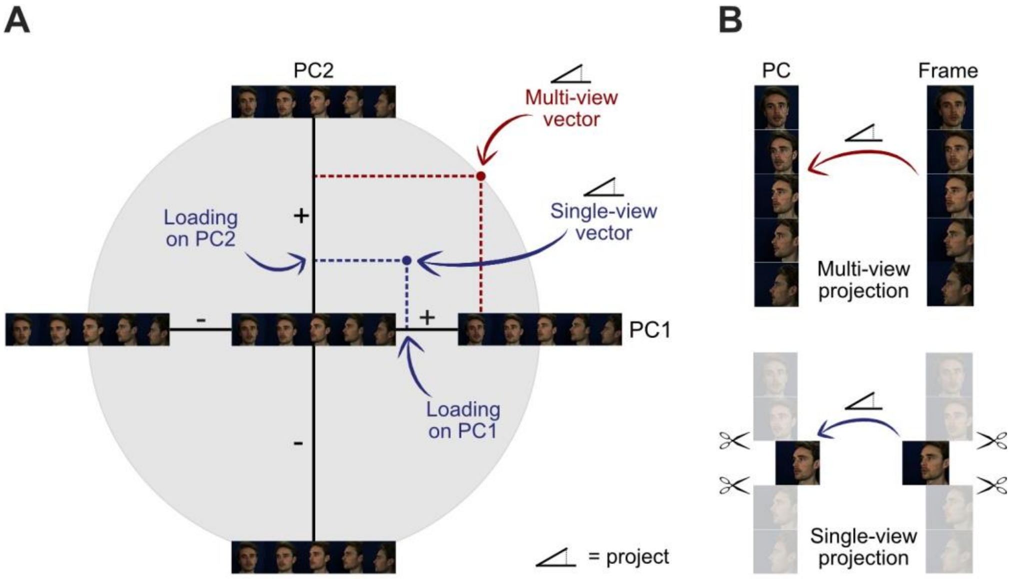


Figure 4.9. Projecting single- and multi-view frames into the multi-view PCA space.
(**A**) Projecting the full multi-view vector and the single view vector into the multi-view PCA space. (**B**) Depiction showing how all views are retained and projected onto each other in the multi-view projection (top) and how the input frames and components are truncated to the input view in the single-view projection (bottom). Images rather than deviation vectors are shown for illustration purposes only.

To reconstruct a frame for all views, the eigenframes were scaled by the frame’s multi-view loadings, summed, and the mean multi-view vector $\bar{X}$ added. The resulting vector was then separated by view and de-warped back into a set of 5 images. Projecting the multi-view vectors into the PCA space acted as a baseline due to creating the most accurate reconstructions possible using projection alone.

The same procedure was performed for single views. Deviation vectors and eigenframes were first truncated to the input view. Initially, inner product loadings were calculated between the truncated vectors; as outlined in section 4.3.2, the outcome is the same as replacing other views' elements with zeros (Beridze, 2021; Scholes et al., 2020). The loadings scaled the complete, multi-view eigenframes to make the reconstructions. A summary of projecting both the full multi-view and single-view vectors into the PCA space can be seen in Figure 4.9.

4.6.4 Scaling

For single-view inputs, the inner product loadings (see section 4.3.2) result in some transfer of expression across views, but the movements are very muted. This was also observed by Beridze (2021) and, to a lesser extent, by Scholes et al (2020). Less muting in Scholes and colleagues' model is likely because they were projecting onto ~½ of the vector whereas we and Beridze had 5 and 6 viewpoints, thus were projecting onto a fifth and a sixth of the vectors respectively. An example of this problem is demonstrated in section 4.3.2.

A few options were considered for scaling single-view loadings, however, as outlined later, there is more to the problem than just scaling. Option 1 is no scaling (Lx). Option 2 multiplied all loadings by a uniform scaling factor of 5 ($5Lx$) due to projecting onto approximately 1/5th of the eigenframes. As the multi-view inputs provided accurate reconstructions the subsequent options utilised the multi-view loadings, with a similar procedure to Beridze (2021). We could have determined scaling factors from the single and multi-view loadings on a frame-by-frame basis, e.g., using equation (4.3), however the aim was to reconstruct one view from another and eventually eliminate the use of the multi-view vectors. Therefore, we wanted to establish reliable scaling factors for each component that could be used for new frames.

In equation (4.3), $L_{multi-view}$ is the loading from projecting a multi-view vector in the space and $L_{single-view}$ from a truncated, single-view vector. λ is the resultant scaling factor.

$$\lambda = \frac{L_{multi-view}}{L_{single-view}} \qquad \textbf{(4.3)}$$

Initially, to determine consistent scaling factors a series of permutations were run on the frames used for training, similar in approach to Beridze (2021). In each of 100 permutations, all 5 single-view vectors and the multi-view vectors for 100 random frames were projected into the space. Small sets of frames were used to prevent overfitting. For each component and view, a linear least squares (LLS) regression (using $Ly \; = \; \beta 1Lx + \; \beta 0$) was performed to predict the multi-view loadings (Ly) from the single-view loadings (Lx, see Figure 4.10). This provided a scaling factor ($\beta 1$) and a translation factor ($\beta 0$), which were remarkably stable across permutations (see Figure 4.11). Therefore, option 3 multiplied the single-view loadings by the median scaling factor for each component for the given view ($\beta 1Lx$). Option 4 also added the median translation factors ($\beta 1Lx + \; \beta 0$).

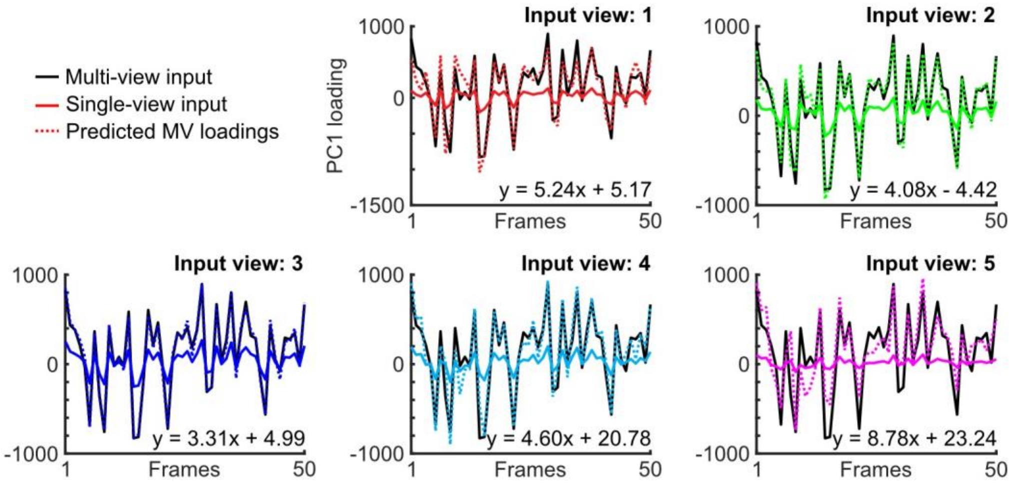


Figure 4.10. Loadings of the single-view and multi-view vectors on PC1 for 50 frames.

Loadings on PC1 of the single-view vectors (solid coloured line) and the full multi-view vector (solid black line) for 50 frames within one permutation. The dashed coloured lines show the predicted multi-view loadings (Ly) from the single-view loadings (Lx) scaled using $Ly = \beta 1 Lx + \beta 0$. Each plot outlines the values for $\beta 1$ and $\beta 0$ from the LLS regression from one permutation. Plots are for one actor.

In general, the scaling factors decreased non-linearly with increasing component number (see Figure 4.11) perhaps expectedly given the hierarchical ordering of the components. Nevertheless, it predicted that a uniform $5x$ scaling factor would overscale most loadings. As the scaling factors decreased exponentially, option 5 scaled the loadings using the predicted values from exponential curve fits (eLx). On each permutation an exponential curve with 3 free parameters was fit to the $\beta 1$ scaling factors using MATLAB's *nlinfit* function. The parameter coefficients were averaged over the permutations producing an average curve for each view (see Figure 4.11).

A simpler method for identifying scaling factors was also assessed (option 6, λLx). All training frames were projected back into the multi-view space, and the median simple scaling factors (λ) were calculated by dividing the multi-view loadings by the single-view loadings, as in equation (4.3).

The eventual aim was to ascertain appropriate scaling without projecting the full multi-view frame back into the PCA space, as this reduced biological plausibility. The multi-view inputs and components are constrained by the correlated facial structure across the views. This constraint means it should be possible to determine consistent scaling factors for each component without projecting in multi-view vectors. This is where the *relative loadings* (see section 4.3.2) come in. When split by view, the truncated components are no longer unit vectors and their lengths differ. As shown in section 4.3.2, assuming the views are correlated, the relative loadings are more stable against truncation and to differences in magnitude across the views compared to inner product loadings. Therefore, option 7 (L_{rel}) used the relative loadings, by dividing the inner products between tD_v and tEF_v by the squared length of the truncated component.

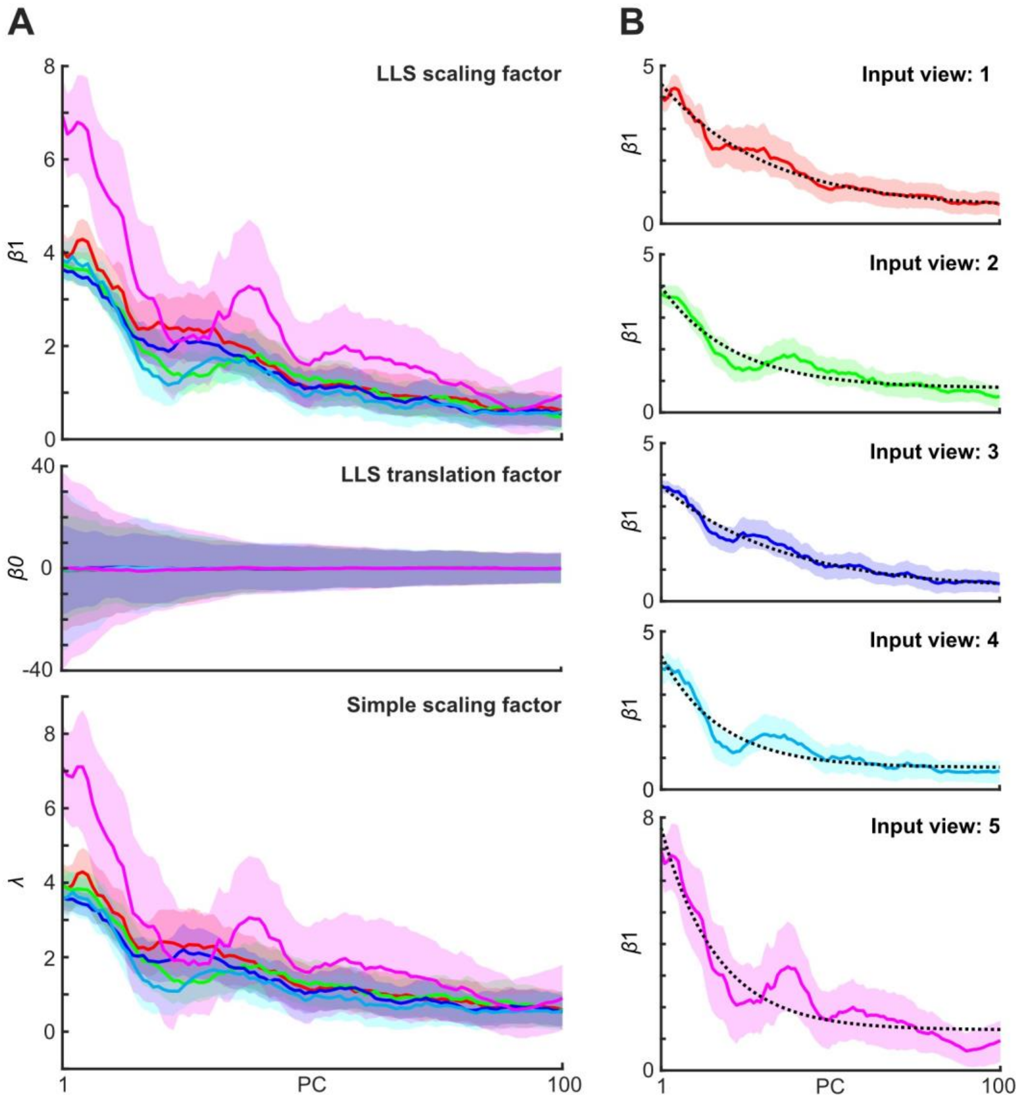


Figure 4.11. Average scaling and translation factors determined over 100 permutations.

(**A**) Average scaling factor (top panel) and translation factor (middle panel) across 100 permutations of LLS regression for each PC showing all views. Average simple scaling factor (bottom panel) from dividing multi-view loadings by the single-view loadings. (**B**) The average scaling factors from LLS separated by view. Solid coloured lines = median, coloured area = 95% range of the average across permutations. Dashed black line = the average exponential curve fit. The data has been smoothed for display purposes.

4.6.5 Comparing scaling methods

To compare different scaling methods, reconstruction accuracy was measured and averaged across training frames for each input and reconstructed view for the single-view inputs (25 values) and each reconstructed view for the multi-view inputs (5 values). The average across the single-view input values was then calculated providing an overall accuracy measurement per actor per scaling method. Two, one-way within-subjects ANOVAs were performed at the group level (across actors) to compare the loading similarities across the 7 scaling options, one using the average Z-value for each actor as the dependent variable and the other the ED. Individual scaling methods were then compared with post hoc t-tests.

4.6.6 Results

4.6.6.1 Principal components

Example components can be seen in Figure 4.12, Figure 4.13 and the supplementary videos. For this, the training frames were projected back onto the components, and the components were scaled by ±3 standard deviations of the loadings. This shows several interesting findings. Firstly, components generally show corresponding changes in all 5 views (Figure 4.12). Secondly, they contain combinations of rigid and non-rigid motion. Thirdly, they often differ across actors (Figure 4.13).

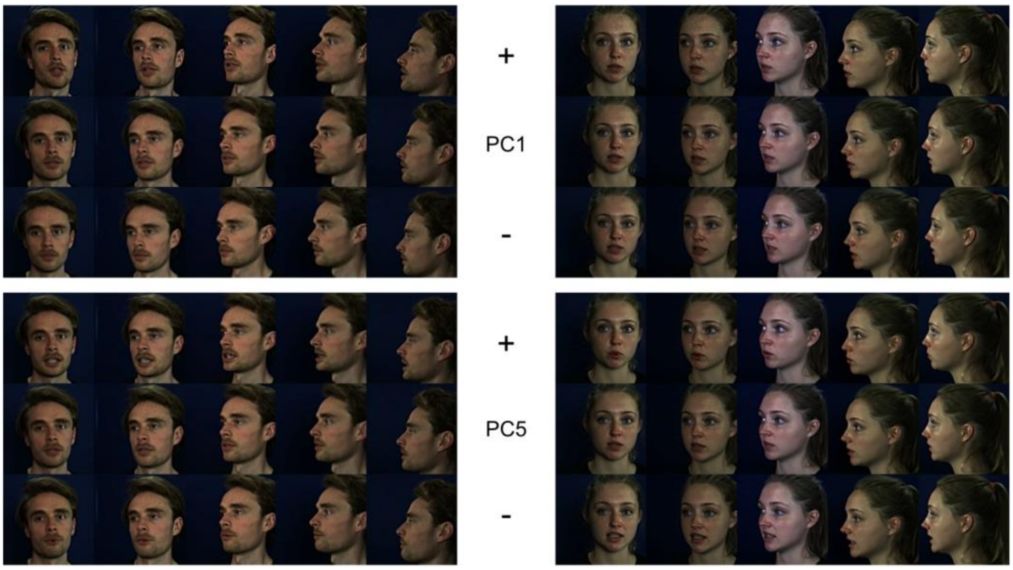


Figure 4.12. Example multi-view components in Model 3.

Depictions of PCs 1 (top) and 5 (bottom) of the multi-view space for two actors. In each set the middle row represents the origin of the multi-view space. The top and bottom rows represent scaling the components by +3SD and -3SD respectively, where SD is the standard deviation of the loadings of the training set on the components.

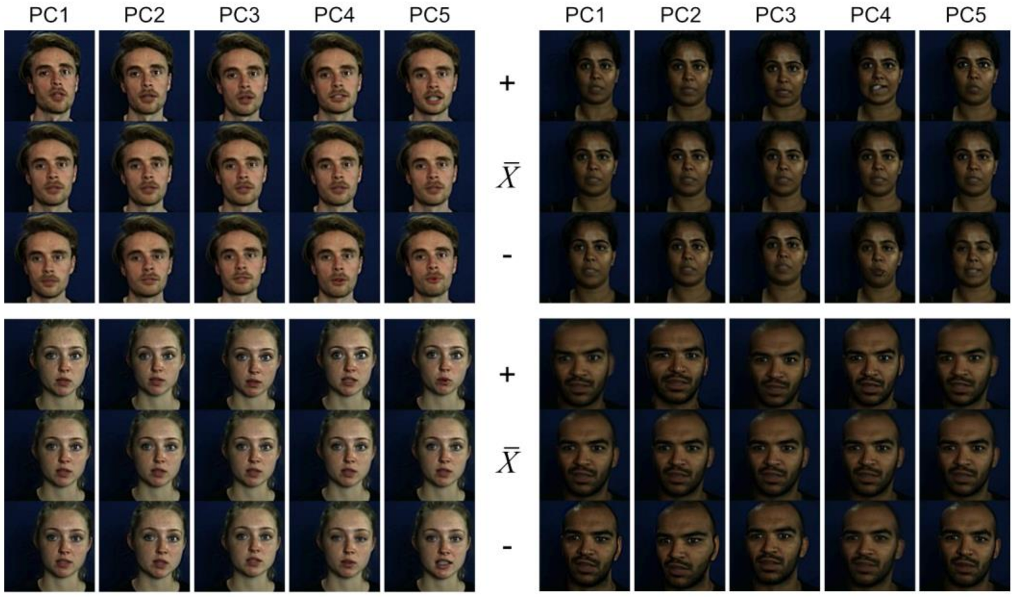


Figure 4.13. Example components across actors in Model 3. Depictions of PCs 1 to 5 of the multi-view space for four actors, showing the front view only. In each of the 4 sets, the middle row represents the origin of the multi-view space. The top and bottom rows represent scaling the components by +3SD and -3SD respectively, where SD is the standard deviation of the loadings of the training set on the components.

4.6.6.2 Reconstructions

Reconstruction accuracy was good for reconstructions made from multi-view inputs. As noted above, reconstructions from the single-view inputs without any scaling were muted. For some scaling options, reconstruction accuracy was substantially improved, with visually accurate reconstructions.

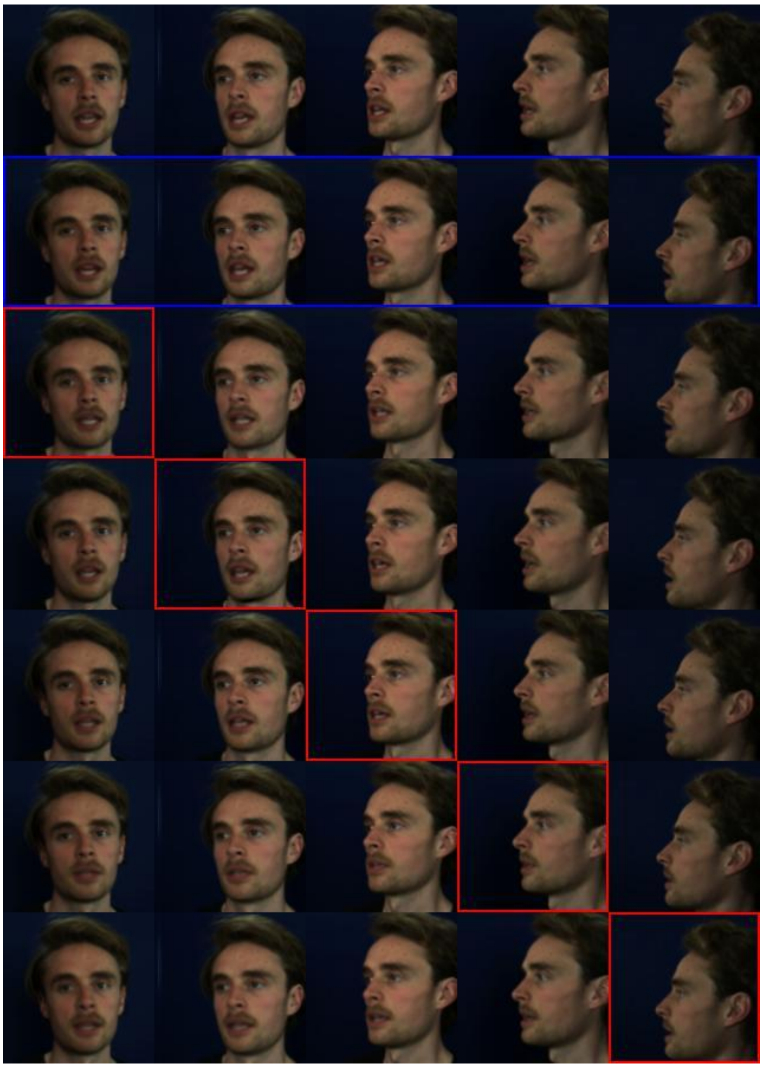


Figure 4.14. Example reconstructions from Model 3 with scaling. Top row shows the veridical frames. Second row (bounded by the blue box) shows the reconstructions made from projecting the multi-view vector into the PCA space. Subsequent rows show the reconstructions made by projecting the single-view vectors (for the views bounded by the red boxes) into the PCA space. This example shows the reconstructions made using the median scaling and translation factors ($\beta 1 L x + \beta 0$).

See supplementary materials for example reconstruction videos as well as Figure 4.14 for one such frame using the median scaling and translation factors (scaling option 3). The average reconstruction accuracy for this model for one participant, separated by input and reconstructed view, is shown in Figure 4.15. Figure 4.16 and Figure 4.17 show a comparison of the scaling methods.

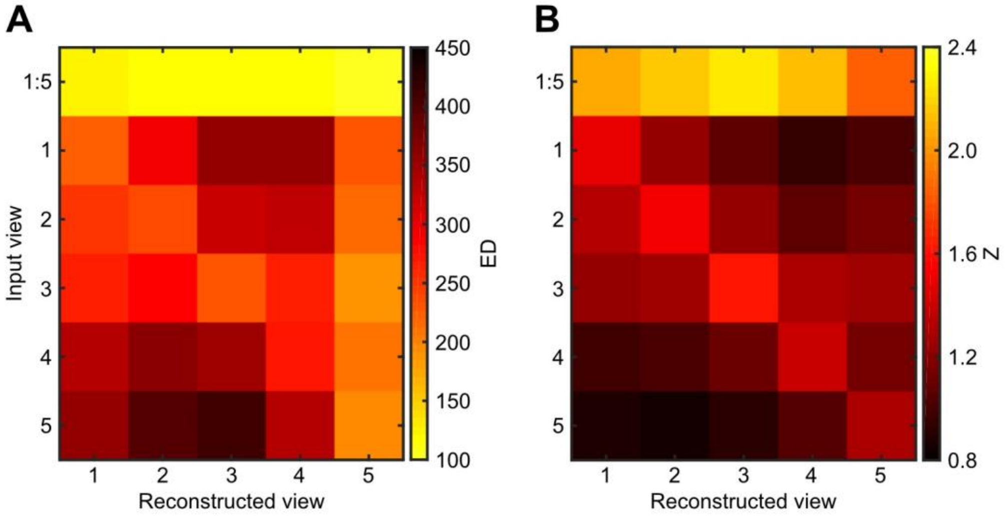


Figure 4.15. Mean loading similarity for one actor, separated by input and reconstructed views.

The mean reconstruction accuracy for each input *x* reconstruction view across all training frames for one actor when '$\beta 1 Lx + \beta 0$' scaling is applied. The heatmaps show the ED (**A**) and the Z (**B**) of the loading similarity. 1:5 = the full multi-view vector.

4.6.6.2.1 Loading similarity: Z

There was a significant main effect of scaling method (F(1.20, 9.61) = 114.84, p < .001, η_p^2 = 0.93, Greenhouse-Geisser correction applied). The Z-values for no scaling and uniform $5x$ scaling were identical, which was expected as a uniform scaling factor should change the magnitude but not the direction of the projection into the separate view space. As such, post-hoc comparisons did not separate these options. The results using the LLS scaling factor and both scaling and translation factors were also so incredibly similar (likely because the median translation factor was close to 0) that post hoc comparisons only included the model using both scaling and translation factors. The Bonferroni corrected alpha for these post hoc comparisons was therefore α = .005.

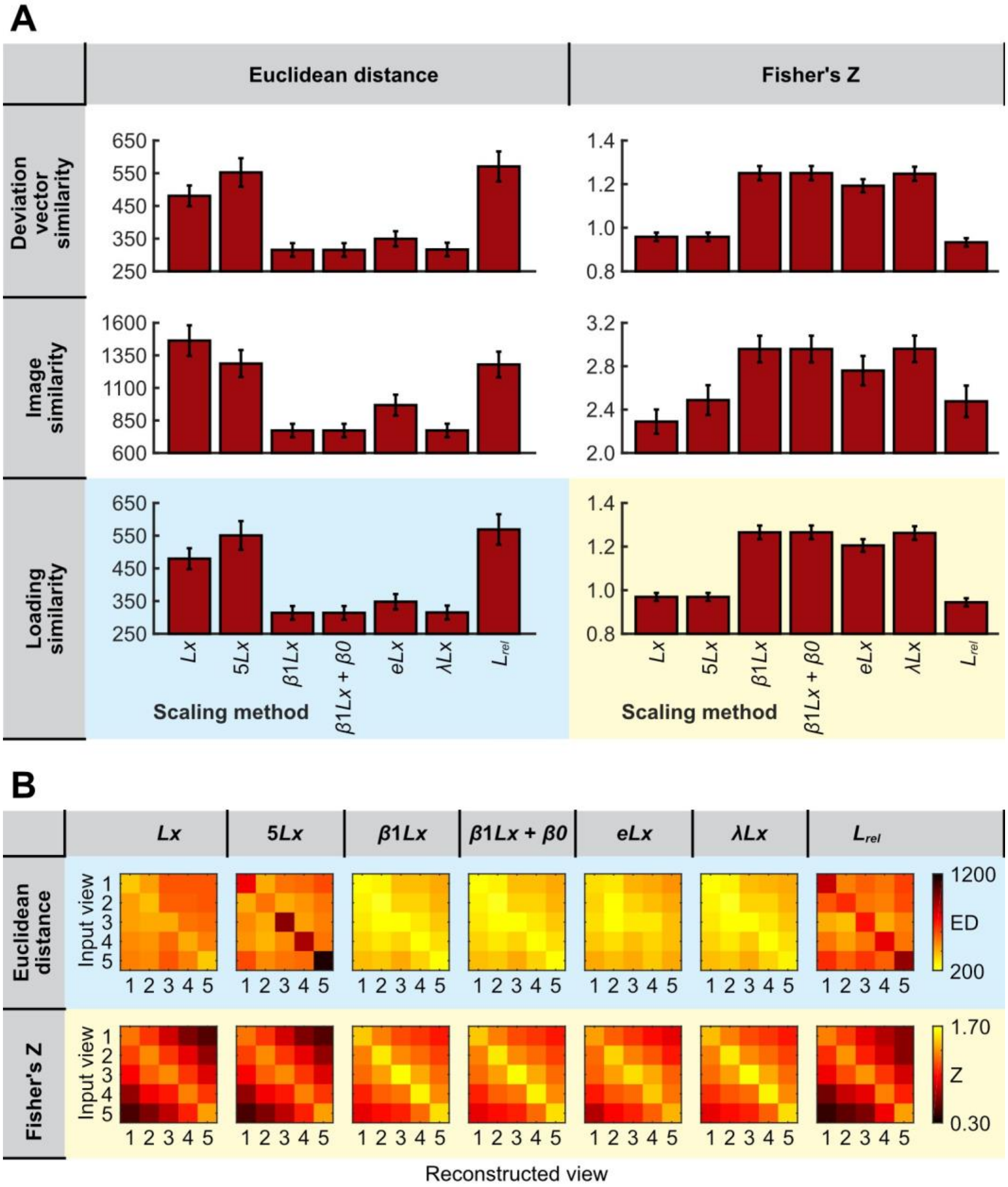


Figure caption on next page

Figure 4.16. Comparison of scaling methods.
(**A**) Mean reconstruction accuracy (ED – left, Z – right, deviation vector similarity – top, image-level similarity – middle, loading similarity - bottom) across actors for each scaling method. 'Lx' = no scaling. '$5Lx$' = uniform $5x$ scaling. '$\beta 1Lx$' = applying the average scaling factor from the LLS regression. '$\beta 1Lx + \beta 0$' = applying the average scaling and translation factors from LLS. 'eLx' = applying the exponential scaling. 'λLx' = applying the simple scaling factor, not from the permutations. 'L_{rel}' = using the relative loadings rather than just the inner products between the truncated input vectors and eigenframes. Error bars show ± 1 SEM of the within-subject mean. (**B**) The loading similarity separated by view for each scaling method, showing the mean ED (top) and Z (bottom) across actors. In all heatmaps, brighter colours reflect better reconstruction accuracy.

In the post hoc comparisons, compared to no scaling (mean = 0.969, SD = 0.055), applying the median scaling and translation factors determined through LLS (mean = 1.265, SD = 0.094) significantly increased the loading similarity Z-value (*t*(8) = 10.60, *p* < .001, 95% CI = [0.232, 0.361]). Applying the median exponential curve fit also increased the loading similarity (mean = 1.205, SD = 0.087) compared to no scaling (*t*(8) = 10.08, *p* < .001, 95% CI = [0.182, 0.290]), however was still significantly worse than using the median scaling and translation factors (*t*(8) = -4.97, *p* = .001, 95% CI = [-0.088, -0.032]).

Using the median simple scaling factor (mean = 1.262, SD = 0.093) was significantly better than no scaling (*t*(8) = 10.80, *p* < .001, 95% CI = [0.231, 0.356]) and exponential scaling (*t*(8) = 4.74, *p* = .001, 95% CI = [0.029, 0.085]) and did not differ significantly from applying scaling and translation factors from LLS ($t(8) = -2.03$, $p = .077$, 95% CI = [-0.006, 0.000]).

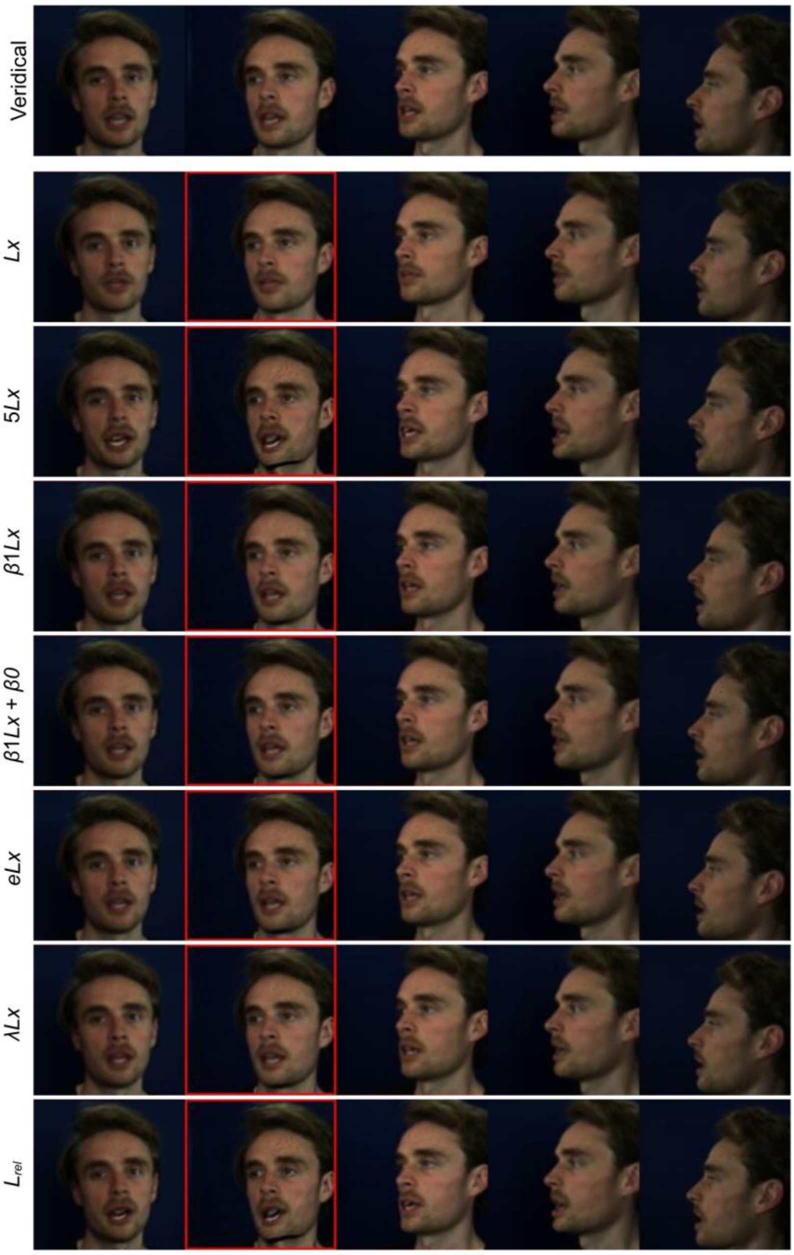


Figure caption on next page

Figure 4.17. Sample reconstructions using different scaling methods.
Sample reconstruction of an untrained frame using view 2 (bounded by red box) as the input view. 'Lx' = no scaling. '$5Lx$' = uniform $5x$ scaling. '$\beta 1Lx$' = applying the average scaling factor from the LLS regression. '$\beta 1Lx + \beta 0$' = applying the average scaling and translation factors from the LLS regression. 'eLx' = applying the exponential scaling. 'λLx' = applying the simple scaling factor, not from the permutations. 'L_{rel}' = using the relative loadings rather than just the inner products between the truncated input vectors and eigenframes.

Surprisingly, using the relative loadings (mean = 0.944, SD = 0.056) rather than inner product loadings was worse than all scaling methods, including no scaling, although this did not survive correction (*t*(8) = -3.64, *p* = .007, 95% CI = [-0.040, -0.009]). This will be returned to later.

4.6.6.2.2 Loading similarity: ED

For Euclidean distance, there was a main effect of scaling method (*F*(1.47, 11.75) = 70.36, *p* < .001, η_p^2 = 0.90, Greenhouse-Geisser correction applied). The values for no scaling and uniform $5x$ scaling were now different, as expected. The scaling and scaling and translation options from LLS were again remarkably similar, and so only the comparisons using both factors are reported (*α* = .0033).

Compared to no scaling (mean = 479.74, SD = 95.45), applying a $5x$ scaling factor (mean = 550.70, SD = 131.68) increased the ED in the loading similarity (*t*(8) = 4.41, *p* = .002, 95% CI = [33.84, 108.09]) indicating worse reconstructions. In contrast, applying the scaling and translation factors from LLS (mean = 313.92, SD = 62.77) decreased the ED, providing better reconstructions compared to no scaling (*t*(8) = -9.51, *p* < .001, 95% CI = [-206.04, -125.59]). Using the exponential curve fit (mean = 347.95, SD = 70.39) decreased the ED compared to no scaling (*t*(8) = -7.59, *p* < .001, 95% CI = [-171.85, -91.73]), but was marginally worse after Bonferroni correction than applying scaling and translation factors (*t*(8) = 4.26, *p* = .003, 95% CI = [15.59, 52.46]).

Applying the simple scaling factor (mean = 315.10, SD = 63.22) also decreased the ED compared to no scaling (*t*(8) = -9.59, *p* < .001, 95% CI = [-204.22, -125.06]) and applying exponential scaling, but only prior to correction (*t*(8) = -3.52, *p* = .008, 95% CI = [-62.19, -12.98]). The ED was slightly but not significantly larger compared to applying the scaling and translation factors from LLS following correction (*t*(8) = 3.57, *p* = .007, 95% CI = [0.42, 1.94]).

Using the relative loadings (mean = 568.97, SD = 139.17) was again worse than no scaling (*t*(8) = 4.09, *p* = .003, 95% CI = [38.97, 139.50]).

There was little visible difference between reconstructions made using the exponential curve fits, both scaling and translation options from LLS or the median simple scaling factor, thus any could be used to visually improve reconstructions.

4.6.6.3 Generalising to new frames

The previous subsection used the trained frames to establish if scaling worked sufficiently. Here we assessed if scaling would generalise to new frames and whether any options overfit the training data, e.g., if the exponential fit generalises better than the median scaling factors. For this comparison, only the following conditions were included: no scaling (Lx), exponential fit from LLS (eLx), scaling and translation factors from LLS ($\beta 1Lx + \beta 0$) and simple scaling factors from the single set (λLx). Two, 2 x 4 ANOVAs were conducted to compare frame set (trained, untrained) and the 4 scaling methods on the ED and Fisher's Z of the loading similarity.

Z: As expected, there was a main effect of scaling method (F(1.22, 9.72) = 74.28, p < .001, η_p^2 = 0.90, Greenhouse-Geisser correction applied). Neither the effect of frame set, nor the interaction were significant.

ED: Again, there was a main effect of scaling method (F(1.49, 11.95) = 80.94, p < .001, η_p^2 = 0.91, Greenhouse-Geisser correction applied). There was also a main effect of frame set (F(1, 8) = 19.16, p = .002, η_p^2 = 0.71), with higher ED for untrained frames (mean = 545.48, SD = 58.25) than trained frames (mean = 364.18, SD = 23.38). The interaction was not significant.

A visual inspection of the loadings in the multi-view space showed they were generally higher for untrained than trained frames.

Despite being quantitatively worse, reconstructions of untrained frames were still visibly accurate. Indeed, the reconstructions presented within the figures are from untrained frames. The lack of an interaction showed all scaling options generalised to new frames comparably.

4.6.6.4 Follow-up analyses on relative loadings

Reconstructions using relative loadings were, contrary to expectations, worse than those using inner product loadings, so we investigated why. Assuming two vectors are correlated, the larger the magnitude of the vector being projected onto, the larger the inner product. As the truncated components vary in length by view, it seemed intuitive that the lengths might inform how to scale the inner products. As such, it seemed intuitive that using the relative loadings $(\frac{tD_v \cdot tEF_v}{|tEF_v|^2})$, would provide suitable scaling compared to simply using the inner product $(tD_v \cdot tEF_v)$.

Further inspection, however, shows the relative loadings are fairly accurate for the first few components, but decrease in suitability as component number increases. As shown in Figure 4.18, simple scaling factors calculated by dividing the multi-view loadings by the relative loadings are now close to 1 for early components, showing that the first few components need little scaling.

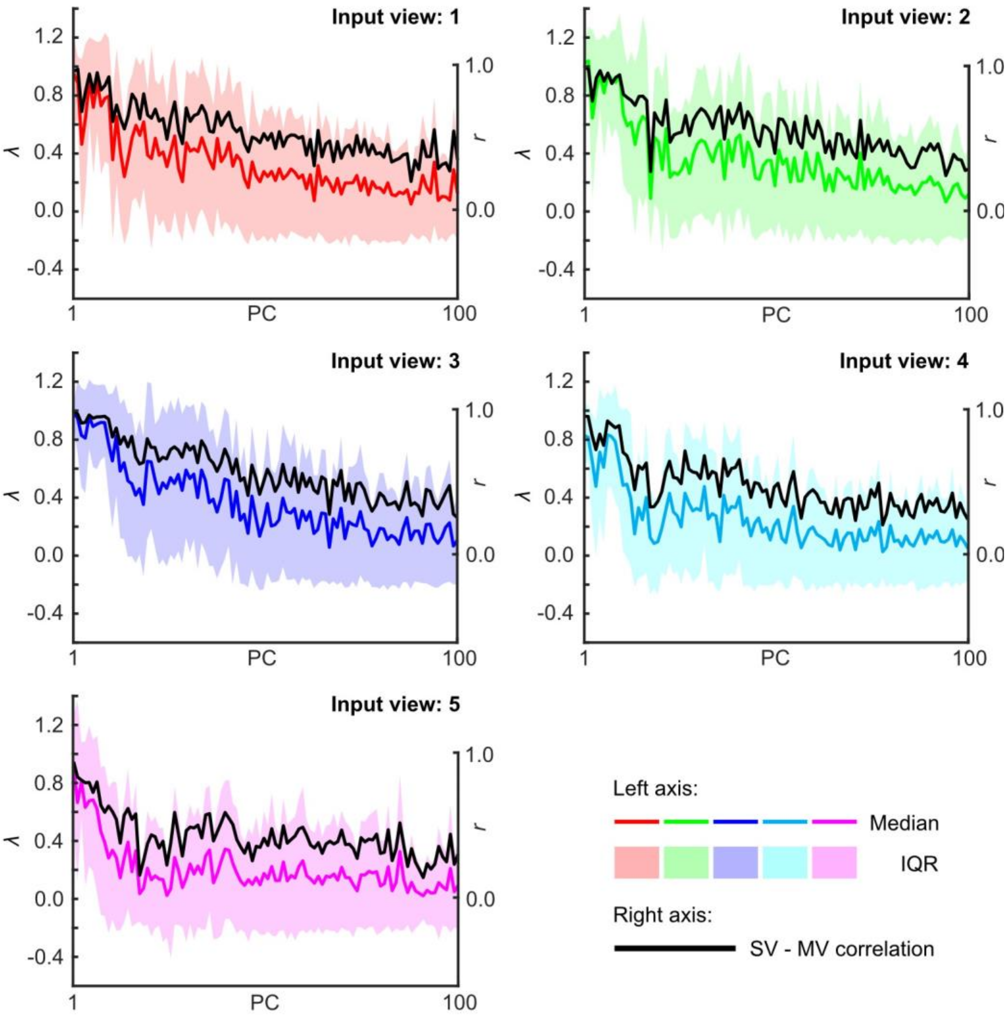


Figure 4.18. Scaling factors using relative loadings against the correlation between single- and multi-view loadings.
The median simple scaling factor (coloured lines) and inter-quartile range (IQR, shaded areas) across all training frames between the single view relative loadings and the multi-view loadings, separated by views. Scaling factors are plotted against the correlation coefficient (*r*) between the single and multi-view loadings across all training frames for each component (black lines). This data has not been smoothed.

We then tested the assumption that single-view and multi-view loadings were correlated; a necessary assumption for scaling to work appropriately. Across all training frames the correlation between the single-view and multi-view loadings was calculated, for each component and view separately. Note that the correlation coefficients are identical regardless of whether inner product or relative loadings are used. These correlation coefficients were plotted against the simple scaling factors for the relative loadings (Figure 4.18), which conveniently sit within a similar numerical range making for easier visual comparison.

As can be seen in Figure 4.18, the patterns of the correlation coefficients and the average simple scaling factors are remarkably similar across components. Indeed, performing a Pearson's correlation between these traces showed a highly significant, positive relationship for all views and actors (all $r > 0.92$, $p < .001$). This shows that, in general, as the component number increases the correspondence between the single-view and multi-view loadings decreases. This decreasing correspondence is reflected in the required scaling factors. The relative lengths of the truncated components across views do not become more variable as the component number increases, suggesting instead the motion and textures in the truncated components become less and less correlated across views. Despite this, it was not clear from observing videos of the components that the actions differed across the views even in higher components.

One approach might be to use relative loadings but only include early components. Figure 4.19 shows how the reconstruction accuracy for one actor changes as a function of either adding more components (red) or removing early components (blue). It shows that reconstruction accuracy decreases rapidly as early components are removed and that it peaks when using the first 17 components. Figure 4.20 shows the reconstruction using the first 17 components.

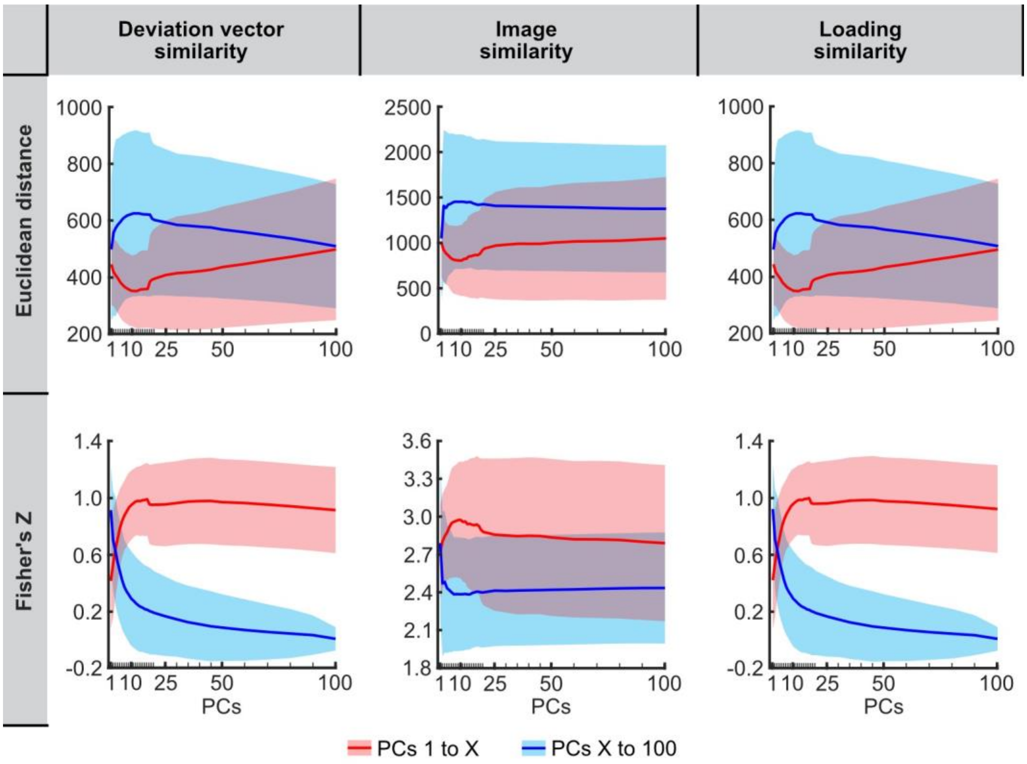


Figure 4.19. Reconstruction accuracy using the relative loadings as a function of the number of components.

Euclidean distance (top row) and Fisher's Z (bottom row) for the deviation vector similarity (left), frame similarity (middle) and loading similarity (right) as a function of the number of components using the relative loadings. Lines show the mean reconstruction accuracy for one actor as a function of either increasing (red) or decreasing (blue) the number of components. Increasing components include PCs 1 to X, decreasing include PCs X to 100. Mean (line) and ±1SD (shaded areas) averaged across all frames and all viewpoints used in training. Combinations of components tested are marked by ticks on the x-axis.

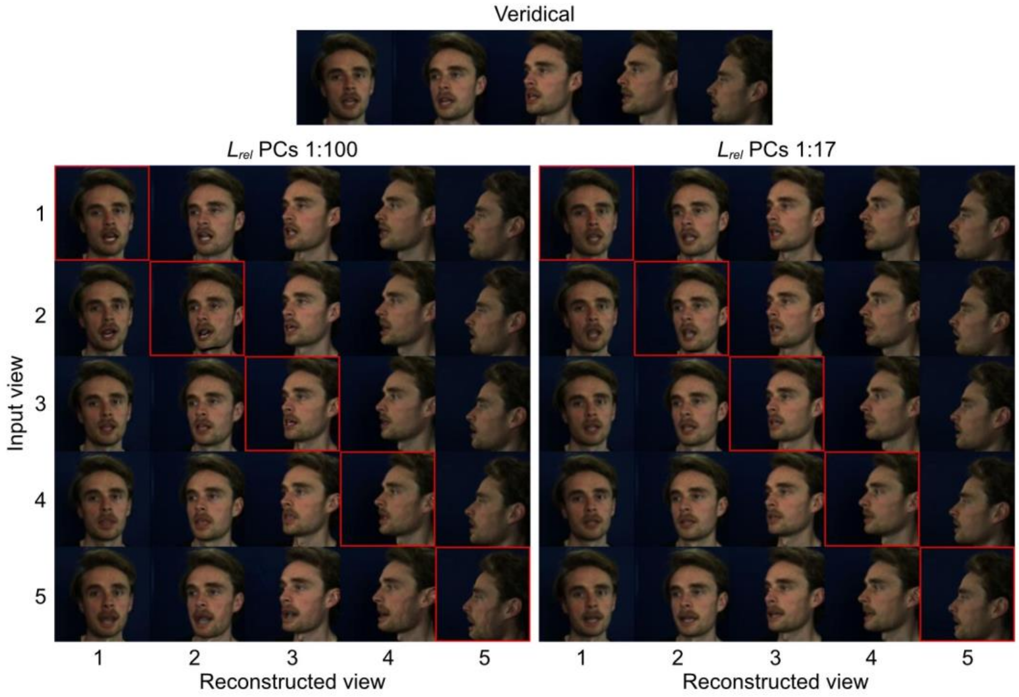


Figure 4.20. Reconstructions using relative loadings. Reconstructions of an untrained frame using relative loadings, using the first 100 components (left) and only the first 17 components (right). The red box outlines the input view for each row of reconstructions.

To test if using only early components improved the reconstructions, the first c components (where $c \leq 20$) were selected for each actor separately to provide the best reconstruction accuracy. This varied substantially across actors, peaking at the first 5 components for one actor and the first 20 for another. Using the first c components significantly lowered the ED of the loading similarity (mean = 407.02, SD = 101.34) and increased the Z-value (mean = 1.063, SD = 0.099) compared to using relative loadings with all components (ED: mean = 568.97, SD = 139.17, *t*(8) = -6.62, *p* < .001, 95% CI = [-218.38, -105.53]; Z: mean = 0.944, SD = 0.056, *t*(8) = 4.52, *p* = .002, 95% CI = [0.058, 0.179]) and compared to using inner product loadings with no scaling, prior to Bonferroni correction (ED: mean =479.74, SD = 95.45, *t*(8) = -2.78, *p* = .024, 95% CI = [-133.07, -12.37]; Z: mean = 0.969, SD = 0.055, *t*(8) = 3.31, *p* = .011, 95% CI = [0.029, 0.160]). Bonferroni-corrected α = .008. The reconstructions were still significantly worse than using the median scaling and translation factors ($\beta 1Lx + \beta 0$) from the linear least squares scaling (ED: mean =313.92, SD = 62.77, *t*(8) = 5.05, *p* = .001, 95% CI = [50.51, 135.68]; Z: mean = 1.265, SD = 0.094, *t*(8) = -18.12, *p* < .001, 95% CI = [-0.228, -0.176]).

Despite cropping the number of components to obtain the maximum reconstruction accuracy when using the relative loadings, the reconstructions were still inferior to those made by scaling the inner product loadings.

4.6.7 Discussion

Here we aimed to recreate the multiple appearance subspace model created by Beridze (2021). As well as recreating the methods, we explored additional options for scaling the reconstructions. Model 3 provided visibly accurate cross-view reconstructions when scaled using the exponential curve fit, the median scaling and optionally translation factors from LLS, and the median simple scaling factor when using the inner products. Quantitatively, the best reconstructions used the scaling factors from LLS permutations, narrowly followed by the median simple scaling factors taken over all frames. Unlike Models 1 & 2, Model 3 did not result in any unwanted superimpositions or suffer from any problems of projecting features in the wrong spatial location.

The model could reconstruct facial motion across views, but the methods were not biologically plausible. In many instances the model was exposed to multiple views concurrently. One such instance is in the projection of the multi-view frames back into the space for determining scaling factors. While the scaling factor curves loosely reflected the decrease in variance explained by each component, they could not be fully explained by the variance as the plotted scaling factor curves would be smoother and identical across views.

By splitting the components by view it was possible to determine if the required scaling factors were in some way related to the lengths of the truncated components. However, reconstructions using the relative loadings, which scaled the inner products by the squared length of the truncated vectors, were poor even when the number of components was specifically cropped to produce the best reconstruction accuracy.

Exploring the correlation between the single-view and multi-view loadings revealed that the views become less and less correlated with each other as the component number increased. The high similarity of the pattern of correlation coefficients across components compared to the required scaling factors suggests that the scaling factors are a product of this decreasing correspondence. As a result, the scaling factors subsequently downregulate the loadings in higher components.

Overall, we managed to replicate the findings by Beridze (2021) showing that facial motion could be reconstructed across views, but again showing the necessity to scale the otherwise muted reconstructions. Models 4 and 5 expand on this model to enhance the biological plausibility and overcome methodological limitations, paving the way to understanding how a face space account might allow for translations across view.

4.7 Model 4 – Multi-view paired entry model

4.7.1 Model aim and overview

Model 3 was able to reconstruct motion across views well once scaling was established. Model 4 aimed to build a similar multi-view space, but without being exposed to all 5 views simultaneously. Based on the assumption of broad tuning curves (Perrett et al., 1991), a single view likely activates at least 2 sets of neighbouring view-tuned neurons. Therefore, we only concatenated frames for neighbouring views. If the head then turns, we remove the first view, keep the second, and add a third, providing overlap between viewpoints. This can be visualised in Figure 4.21.

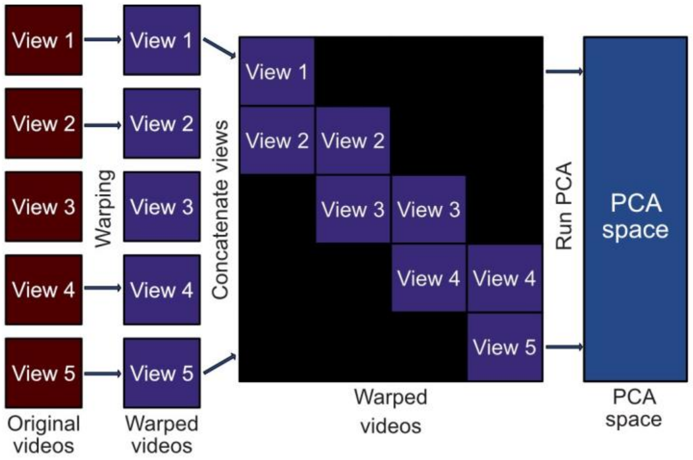


Figure 4.21. Depiction of Model 4.

The multi-view paired entry model. Vectorised versions of the frames are concatenated across views in neighbouring pairs, with the remaining views filled with zeros, before performing PCA on this multi-view matrix.

4.7.2 Model construction

As with Model 2 (the single slot model), the mean for each view ($\bar{X}_v$) was subtracted prior to concatenation. The current model however followed the multi-slot design of Model 3. Deviation vectors for two neighbouring views were concatenated, with the rest of the vector filled with zeros. The matrix X was therefore of size 409,600 x $4n$. Different pairs would realistically be seen at different times, but for piloting, the same set of n frames were reused for each pair. PCA was performed on X, with the hope that through the overlap, correlated changes across all 5 views would be learned.

4.7.3 Reconstructing frames across views

Reconstructing frames across views used the same methods as Model 3, although, less time was invested in scaling because of flaws discovered. In some components, actions were reversed for some views. For instance, in some views the head leaned forwards for positive loadings and backwards for negative, while in others the action was reversed, leaning backwards for positive loadings. The polarity in general is arbitrary, with opposing directions simply reflecting opposite changes from the mean, but one would have hoped the directions of actions would be consistent within components.

To attempt to fix the polarity inversion, a chain of comparisons was performed, starting with views 1-2, then 2-3, and so forth. In each comparison, frames for both views were projected onto their respective truncated components. Where the sign of the loadings was flipped between views, those components in the second view were multiplied by -1. This *should* have corrected the polarity across views and the direction of actions *should* then have matched. As shown in the results this was not the case, and more issues were present.

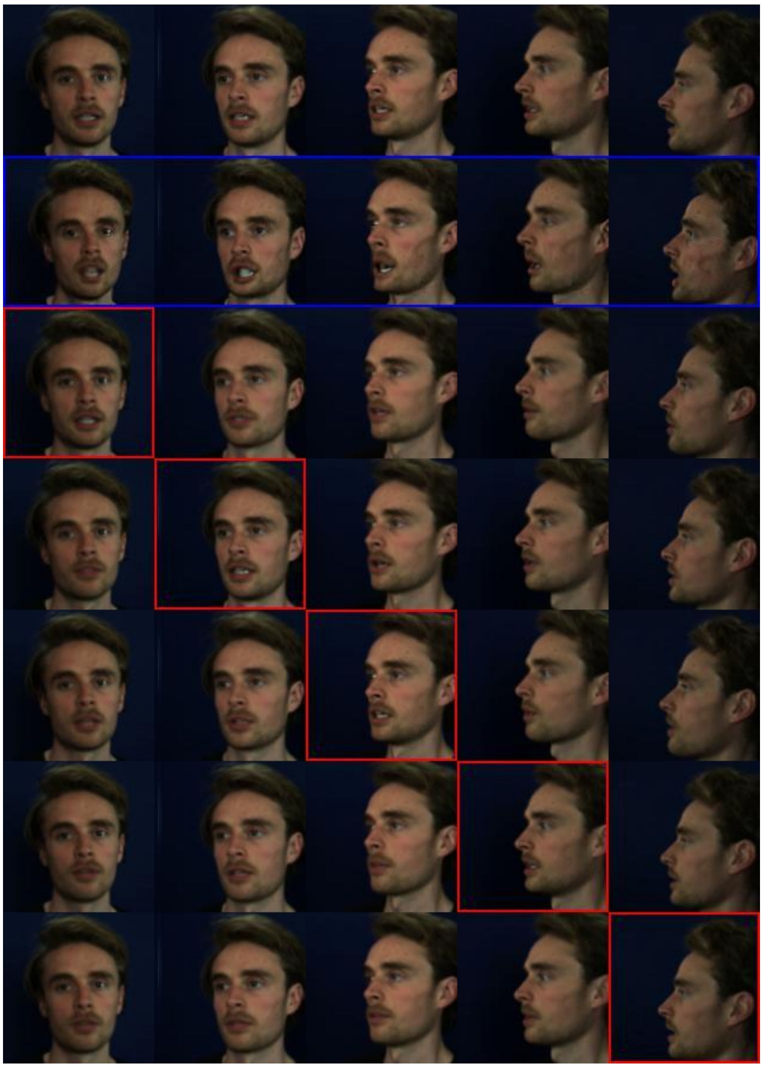


Figure 4.22. Example reconstructions from the multi-view, paired entry model.

Top row shows the veridical frames. Second row (bounded by the blue box) shows the reconstructions made from the multi-view input. Subsequent rows show the reconstructions made by projecting the single-view vectors (for the views bounded by the red boxes) into the space. These reconstructions were made using the inner product.

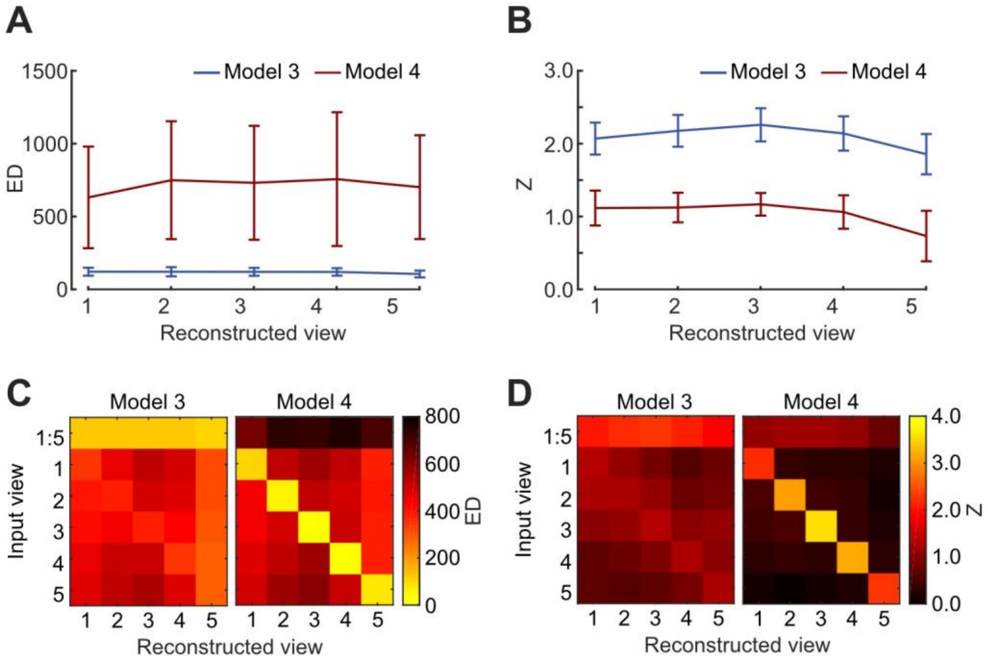


Figure 4.23. Comparison of Models 3 and 4.

(**A&B**) The mean Euclidean distance (**A**) and Fisher's Z (**B**) of the loading similarity for reconstructions from the multi-view inputs for Model 3 (blue) and Model 4 (red) for one actor, separated by reconstructed view. Error bars show ±1SD across all trained frames. (**C&D**) The mean Euclidean distance (**C**) and Fisher's Z (**D**) separated by both input and reconstructed view. The top row in each heatmap shows the reconstructions from the multi-view inputs, the subsequent rows the single-view inputs.

4.7.4 Results and discussion

In Model 3, reconstructions from multi-view inputs were better than from single-view inputs, providing a baseline of what was achievable and providing loadings that could help determine scaling factors. Here, however, reconstructions from multi-view inputs were inaccurate and often caricatured (see Figure 4.22). Figure 4.23 shows that the average reconstruction accuracy (separate space loading similarity) for multi-view inputs was substantially worse than in Model 3.

These reconstructions no longer provide a good baseline, nor can the loadings be used for scaling.

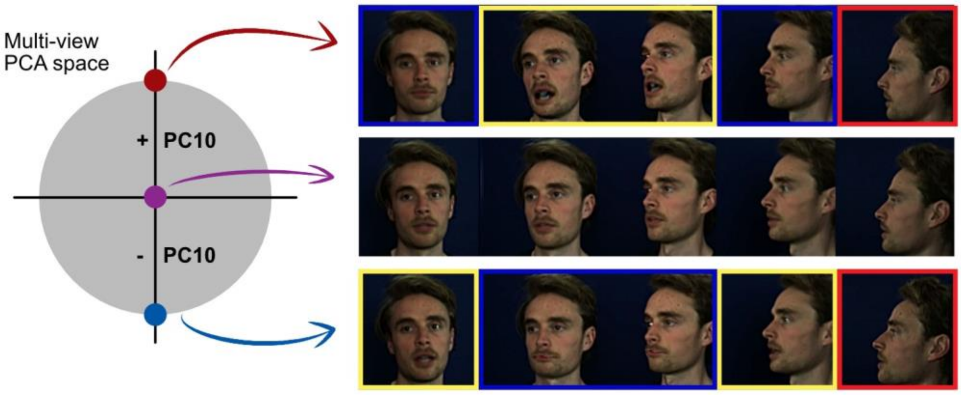


Figure 4.24. Mixed component in Model 4.
Depiction a positive (top) and negative (bottom) loading on PC10, and the origin (middle). Views 1 and 4 show the same action. Views 2 and 3 show the same action but in the opposite direction. View 5 shows a slightly different movement as the lip separation does not change.

Inspecting the components using videos that sinusoidally scaled through the origin and along both polarities revealed three problems at least. Problem 1: within components the amount of motion differed across views, with the actions in some views (e.g., ¾) extending further than others (e.g., frontal). Problem 2: inverted actions were still present, despite loadings being of the same polarity. Problem 3: more problematically, actions did not always perfectly match across views, irrespective of polarity. PC10 (Figure 4.24) of one actor demonstrates this. Views 1 and 4 show the same action, which is the same but reversed in views 2 and 3. View 5 shows some similarities in overall head motion, but the lip separation does not change. This sort of discrepancy may be present in Model 3, explaining the decreasing correlation between single-view and multi-view loadings as component number increased, although, as already noted we did not see any obvious discrepancies like this.

These problems mean the multi-view loadings are a conglomerate of inconsistencies, accumulating in inaccurate and often caricatured reconstructions. For single-view inputs, same-view reconstructions were surprisingly good, likely from loading most onto components more heavily biased to that view. Between differences in the magnitudes of actions within the truncated components, and differences in the actions themselves, the resulting reconstructions were more accurate for the input view, but less accurate for other views.

To summarise, Model 4 was theoretically more biologically plausible than Model 3 while having view-invariant units. However, reconstructions from multi-view inputs were worse due to an even greater lack of correspondence between the separate views within the components.

4.8 Model 5 – Two-step multi-view face space model

4.8.1 Model aim and overview

This final, two-step model accumulates many of the ideas discussed previously. Of the models created, this model provides the best reconstruction accuracy whilst being the most biologically plausible and can account for more observations in the known hierarchy of face processing.

4.8.2 Model construction

An outline can be seen in Figure 4.25. The model comprises two steps: step 1 forms separate spaces for each view, which are then used to create the multi-view space in step 2.

4.8.2.1 Step 1:

Separate PCA spaces were made for each view (example components can be seen in the results section). This followed the same procedure as Model 1, with the assumption of a view-dependent layer containing separate populations of neurons for each view.

Rather than projecting deviation vectors across the spaces, we instead learned how components in the different spaces were associated. This is a similar idea to the canonical correlation models described by Ding and Tao (2016), without explicitly aligning the separate spaces into one common subspace. Instead, we stored each space separately, alongside transformation matrices detailing how to move between them. There are many forms that these transformation matrices could take, such as correlation, projection, or regression matrices.

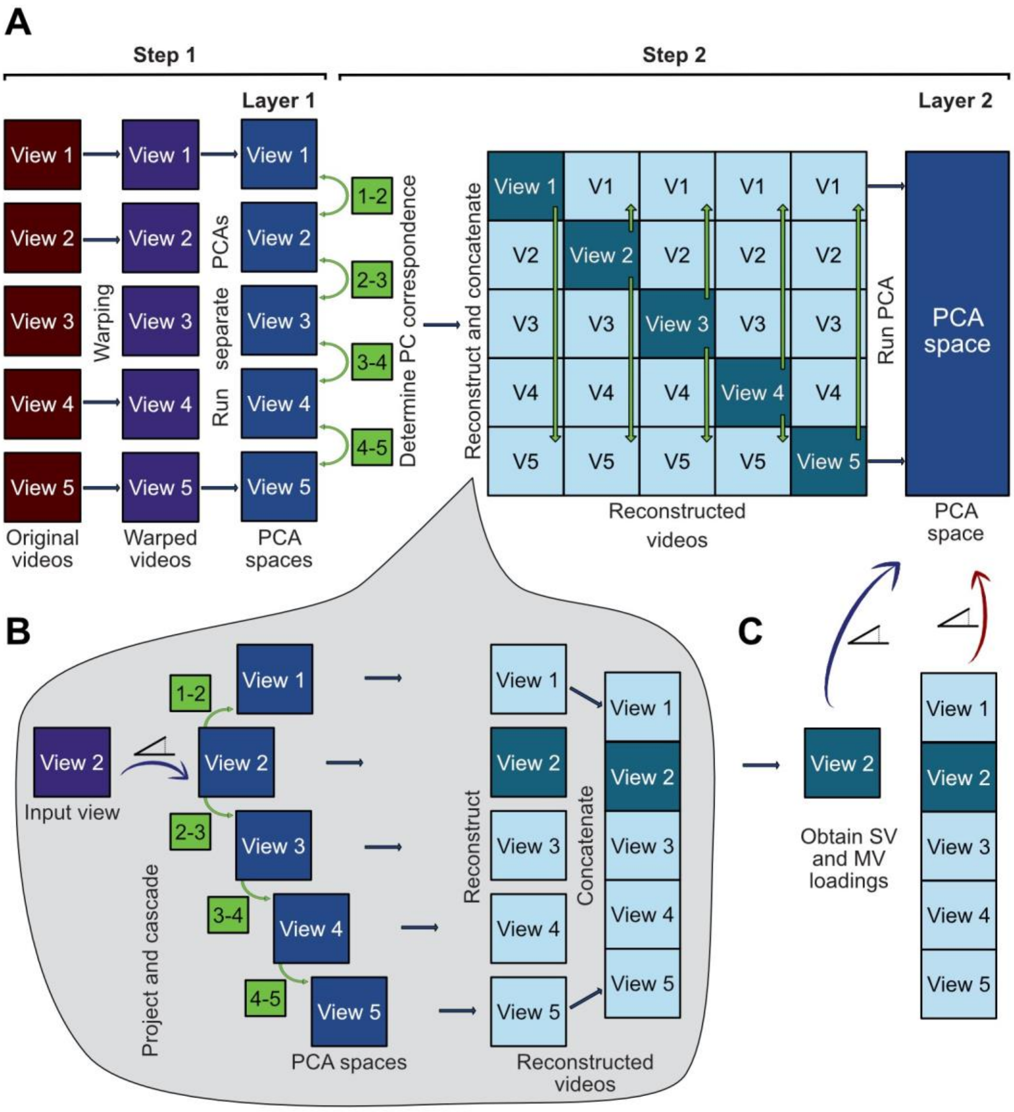


Figure 4.25. Depiction of Model 5.

Depiction of two-step model construction. (**A**) An overview of the steps, showing separate, view specific PCA spaces in Layer 1, and the multi-view space in Layer 2. (**B**) The cascade reconstruction process to recover multi-view images from single-view inputs across the separate spaces. This process uses matrixes of regression coefficients (green) for predicting the loadings in one view's space from those of another. (**C**) Projecting both the recovered single-view (SV) vectors and multi-view (MV) vectors into the multi-view space of Layer 2. Multi-view loadings are used for scaling reconstructions.

A similar approach was used by Lan et al (2012) and Lucey et al (2007) who transformed the appearance of the mouth across views for automatic lip-reading systems. They created feature vectors, similar to our PCA components, containing information regarding shape and texture deviations. These features were similar over views so they could map from one view's space to another directly, but to improve the mapping they used ridge regression to calculate transformation matrices. The features in one view (Y) could be predicted those in another (X) using $Y = TX$ where the transformation matrix T is a matrix of β coefficients. The calculation of $T,$ however, involves computing the covariance between X and Y. This only works because of the close spatial correspondence of the features across views due to being restricted to the mouth. With full faces and larger spatial differences, this method becomes less suitable, as discussed in Model 1.

Rather than using the components of two neighbouring spaces (i and j) as the features for regression, we instead used the loadings of frames onto the components. Multiple linear regression (MLR) was also used instead of ridge regression. Ridge regression was used previously (Lan et al., 2012; Lucey et al., 2007), possibly due to being suitable when features are collinear, however because PCA components are orthogonal, the features cannot be collinear and loadings on the components should not be either. We also did not map directly from profile to frontal (as in Lan et al., 2012; Lucey et al., 2007), instead learning the transformation matrices between neighbouring views only, based on the same assumptions as for Model 4. Like Model 1, reconstructions were made through a cascading procedure across neighbouring views.

We projected frames for two neighbouring views into their respective spaces and calculated the loadings onto the components. As the frames were synchronised across views, we could then learn how the loadings in one space related to loadings in another. For instance, the loadings on PC1 in one space may relate to, and therefore be predicted by, the loadings on PCs 2 and 3 in the neighbouring space. For a given view's space (j), we therefore used MLR to learn the β coefficients necessary to predict the loadings on each component from the loadings on all components in the neighbouring space (i). For this, the loadings on each of the components in i were entered as the predictor variables, and the loadings for one component in j were entered as the response variable. This was repeated for every component in j, and then the process was reversed to predict the loadings on the components in i from the loadings in j.

To reconstruct frames across views, the loadings in one view (Lj) were then predicted from the loadings in another (Li) using equation (4.4). In this equation, Li is a $n\ x\ p_i$ matrix of loadings where n is the number of frames and p_i is the number of components in i. $\beta 0$ is a $1\ x\ p_j$ row vector containing a constant for each component in j, where p_i is the number of components. βij is a $p_i\ x\ p_j$ transformation matrix for transforming the loadings in Li into Lj. Lj is a $n\ x\ p_j$ matrix of the predicted loadings onto j. The components could then be scaled by the loadings and the frames reconstructed.

$$Lj = \beta 0 + LiBij \quad \textbf{(4.4)}$$

To reconstruct non-neighbouring views, the process was cascaded along. As an example, from the loadings in view 2's space (the input view), the loadings in the neighbouring views (1 and 3) were estimated and the frames reconstructed. View 3's estimated loadings were then used to estimate view 4 and in turn view 5. This can be seen in Figure 4.25.

Regression is not the only method for making transformation matrices. Correlation matrices could be used, but from trialling both methods, regression creates more accurate reconstructions. This is noticeable across large rotations, although not strikingly obvious. Because correlation coefficients are limited to ±1, the loadings can only stay the same or decrease when transformed across spaces, muting the motion. Due to its predictive nature, and not being capped at ±1, regression provides a better alternative.

At this point, the frames could be reconstructed remarkably well (see results). As a system, layer/step 1 is view-invariant, but at this stage no single unit is view-invariant. Therefore, step 2 formed a multi-view space as in Model 3.

4.8.2.2 Step 2.

The second step integrated information across views to build a multi-view matrix which was then orthogonalized using PCA forming the multi-view space. It is possible to create this multi-view layer directly from the separate spaces, however here we used the cross-view reconstructions, as if learning through visual input.

Deviation vectors for the training frames were projected back into their respective spaces. The cascade of reconstructions for other views were then calculated and concatenated into multi-view vectors, with each view used as a separate input. This generated a 409,600 x $5n$ matrix where n is the number of frames. PCA was then performed to create the multi-view space.

When generating the reconstructions to form the multi-view space non-significant predictors were ignored. The transformation matrices can be thought of as synapses between neurons in different spaces, or even separate neurons connecting two populations. In either case, if two components do not coactivate often, the connection will be pruned, if ever formed. To mimic this, if any component in one view did not significantly predict loadings on a component in another, then the coefficient for those components was replaced with 0. There is no computational benefit to this in matrix multiplication, but it symbolises one less synapse and fewer neural computations.

There was a significant decrease across all measures of reconstruction accuracy when non-significant predictors were ignored, but the difference was often negligible. For instance, when comparing loading similarity (ignoring same-view reconstructions) the average ED was 220.53 (SD = 45.58) without thresholding, and 220.90 (SD = 45.69) when restricted to only significant predictors - a difference significant at group-level (*t*(8) = 7.88, *p* < .001, 95% CI = [-0.48, -0.26]). The correlation coefficients (Fisher's Z) were on average 1.635 (SD = 0.129) and 1.633 (SD = 0.129) respectively (*t*(8) = 13.40, *p* < .001, 95% CI = [0.001, 0.002]). While significant, the neurocomputational benefit arguably outweighs the negligible difference. There are also no clear visible differences in the reconstructions, therefore non-significant predictor variables were ignored.

One could reduce the number of components in the separate spaces before making the multi-view layer, say to the first 15 as these seem to be crucial for image reconstruction (see Figure 4.26). However, as the reconstruction accuracy of the layer 1 continued to increase with additional components, albeit only slightly, all components were retained to maximise the accuracy of the multi-view layer.

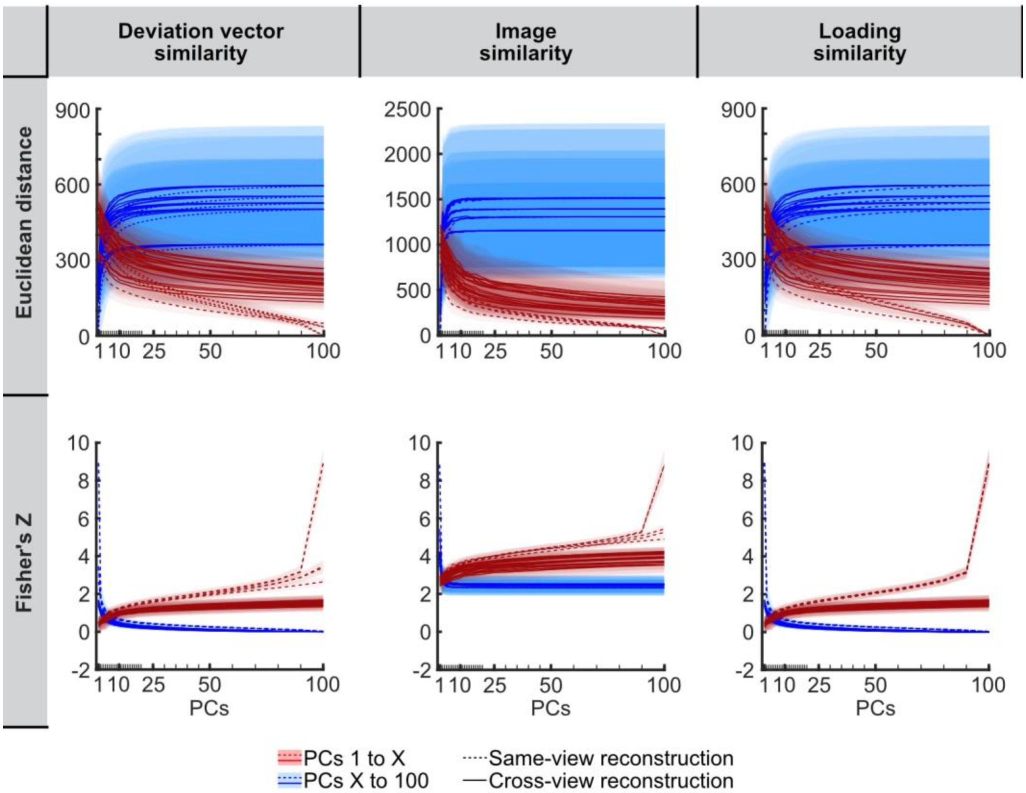


Figure 4.26. Reconstruction accuracy with increasing components in the first layer of the separate space model. Euclidean distance (top row) and Fisher's Z (bottom row) for the deviation vector similarity (left column), frame similarity (middle column) and loading similarity (right column). Lines show the mean reconstruction accuracy for each input x reconstructed view for one participant as a function of either increasing (red) or decreasing (blue) the number of components. Increasing components include PCs 1 to X, decreasing components include PCs X to 100. Mean and SD (shaded areas) for same-view reconstructions (dashed line) and cross-view reconstructions (solid line), averaged across all training frames. The spaces used for measuring reconstruction accuracy contained all 100 components, hence, when all PCs were included, the loading similarity Z-value was inevitably infinite for same view-reconstructions. Infinite values were replaced with 9.557. Not all possible combination of components were tested. Combinations tested are marked by ticks on the x-axis.

The model now contains a 2-layer hierarchy of spaces. The first, view-dependent layer comprises of separate spaces for each view. The second, view-invariant layer comprises a multi-view space where each component has information about all five views.

4.8.3 Reconstructing frames across views

There were multiple options for reconstructing frames across views. One was using the separate spaces and transformation matrices, ignoring the second, multi-view layer. Alternatively, the first layer could be skipped, and the frames directly projected into the second layer. Given the move from view-dependence to view-independence (Freiwald & Tsao, 2010; Meyers et al., 2015), the option best imitating neural processes was to perform a two-step reconstruction process. First, frames for a given view were projected into the corresponding space in layer 1, and deviation vectors reconstructed in that view only. The reconstructed deviation vectors were then projected into the multi-view space as in Model 3 and the remaining views reconstructed.

4.8.4 Scaling

Projecting into the multi-view layer incurred the same problems as in Model 3, with muted reconstructions and therefore the necessity to scale the loadings. The model aimed to improve bio-plausibility, so a modified strategy was adopted. Rather than using all 5 veridical views as the multi-view input for determining the required scaling factors, the cross-view reconstructions from layer 1 were used instead. For instance, all views were reconstructed from view 2 through the cascading process, the multi-view reconstructions were then projected into the multi-view space providing a unique comparison and scaling factors for view 2.

Here, no permutations were run, the median simple scaling factor (λ) and the exponential fit of λ ($e\lambda$) was calculated from all trained frames in one batch. Figure 4.27 shows a comparison of λ derived this way compared to using the veridical multi-view vectors for one actor.

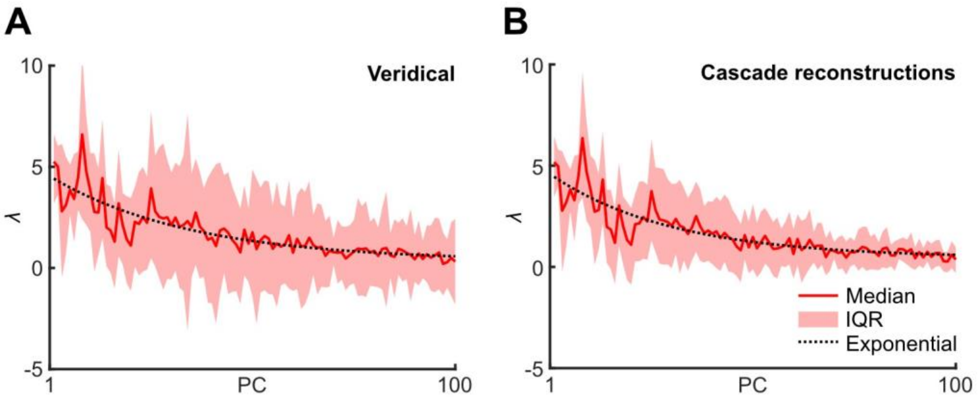


Figure 4.27 Comparison of scaling factors from veridical and reconstructed multi-view vectors.

The required simple scaling factors (λ) for the first 100 components using the veridical multi-view vectors (**A**) and the multi-view reconstructions made from transforming the loadings across spaces in layer 1 of Model 5 (**B**). The solid red line shows the median scaling factor across all training frames for view 1 of one actor. The shaded area shows the inter-quartile range. The black dotted line shows the exponential fit for the data. The median scaling factors are remarkably similar.

4.8.5 Results

4.8.5.1 Layer 1

4.8.5.1.1 Principal components

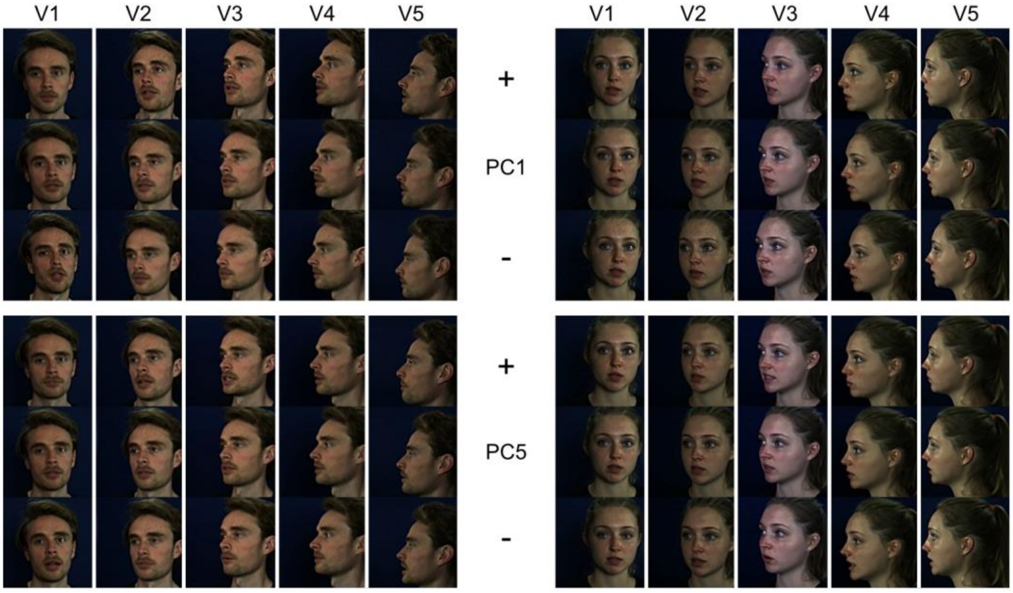


Figure 4.28. Example components in Layer 1 of Model 5.
PCs 1 (top) and 5 (bottom) in the separate, single-view spaces for two actors in the layer 1 of Model 5. In each set the middle row represents the origin of the given view's space. Top and bottom rows represent scaling the components by +3SD and -3SD respectively, where SD is the standard deviation of the loadings of the training set on the components. The different colour for view 3 for the actor on the right is from the auto white balancing from video capture.

Visualising the components (see Figure 4.28 and the supplementary videos) showed that the actions do not necessarily match across different views. As noted for Model 1, this was expected because different viewpoints have access to different spatial and textural information. For example, PC5 for the actor on the right in Figure 4.28 is primarily driven by gaze in view 1's space, and an orofacial change from "eee" to an "ooo" action in view 3. Due to the auto white balancing of the cameras, view 3 for that actor captured whiter images with cooler colours. However, this is not a major concern here, as the PCA components and deviation vectors are both coded as deviations from average, so most of this colour difference will be subtracted out with the average.

4.8.5.1.2 Reconstructions

The fidelity of the first layer was assessed as the accuracy translates to that of the second, multi-view layer. The reconstructions of this cascading process of projection, transformation and reconstruction looked remarkably good (see Figure 4.29 and supplementary videos). They did not require any scaling yet were visibly the best reconstructions from any model, including the second layer of this model.

Ignoring same-view reconstructions, the loading similarly was better in these reconstructions than those made using $\beta 1 Lx + \beta 0$ in Model 3. They had a significantly higher Z-value (Model 5 layer 1: mean = 1.633, SD = 0.129; Model 3: mean = 1.189, SD = 0.105; *t*(8) = 23.68, *p* < .001, 95% CI = [0.400, 0.487]) and a significantly lower ED (Model 5 layer 1: mean = 220.90, SD = 45.69; Model 3: mean = 332.18, SD = 68.55; *t*(8) = -11.67, *p* < .001, 95% CI = [-133.27, -89.30]).

The reconstructions were also better than those using the second layer of the current model, even after scaling was applied to the second layer. The loading similarity Z-value was significantly higher (layer 1: mean = 1.633, SD = 0.129; layer 2: mean = 1.202, SD = 0.104; *t*(8) = 21.64, *p* < .001, 95% CI = [0.385, 0.477]) and the ED significantly lower (layer 1: mean = 220.90, SD = 45.69; layer 2: mean = 329.39, SD = 67.87; *t*(8) = -11.43, *p* < .001, 95% CI = [-130.39, -86.61]).

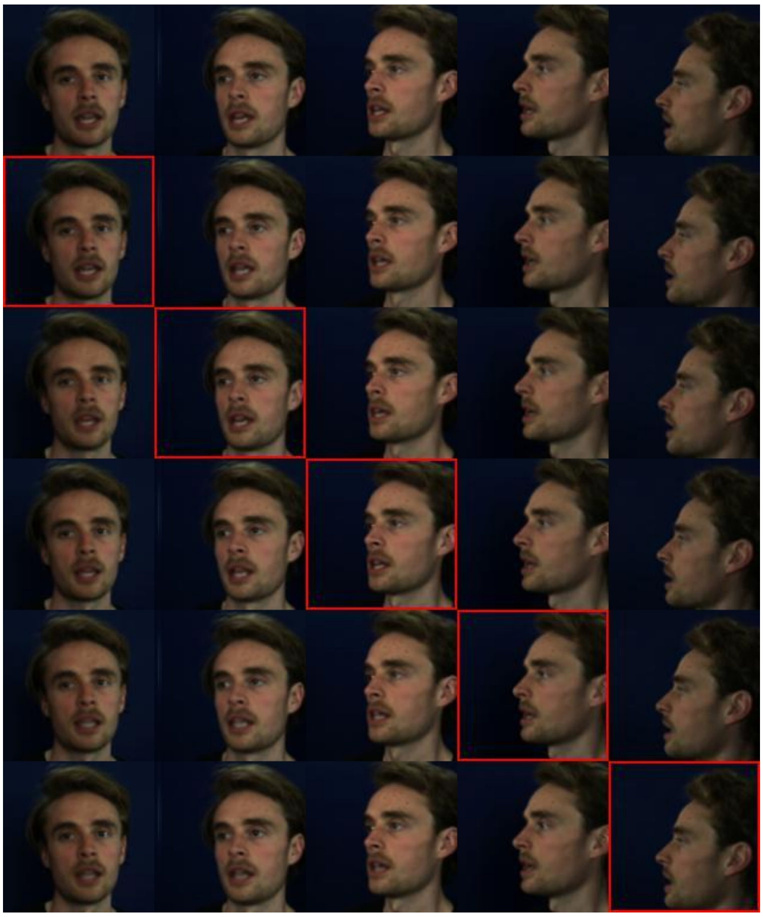


Figure 4.29. Example reconstructions from the first, view-dependent layer of the two-step model.

Top row shows the veridical frames. Each subsequent row shows the cascade of reconstructions made using the view bounded by the red box as the input. Note that there are no multi-view reconstructions when assessing the first layer of the two-step model. The first 100 components of each space were used.

4.8.5.2 Layer 2

4.8.5.2.1 Principal components

Example components for layer 2 can be seen in Figure 4.30 and Figure 4.31. Reassuringly, they depict the same actions as in Model 3 (see section 4.6.6.1). This further demonstrates that the transformations in layer 1 sufficiently captured the correspondence across views for such comparable multi-view components to emerge.

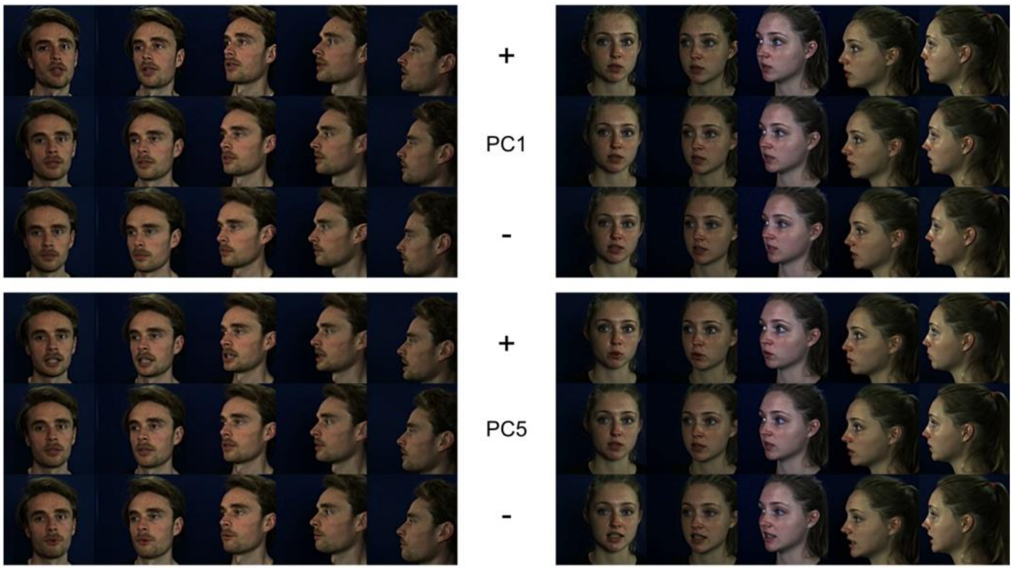


Figure 4.30. Example components in Layer 2 of Model 5.
PCs 1 (top) and 5 (bottom) in the multi-view space for two actors in layer 2 of Model 5. In each set the middle row represents the origin of the multi-view space. The top and bottom rows represent scaling the components by +3SD and -3SD respectively, where SD is the standard deviation of the loadings of the training set on the components.

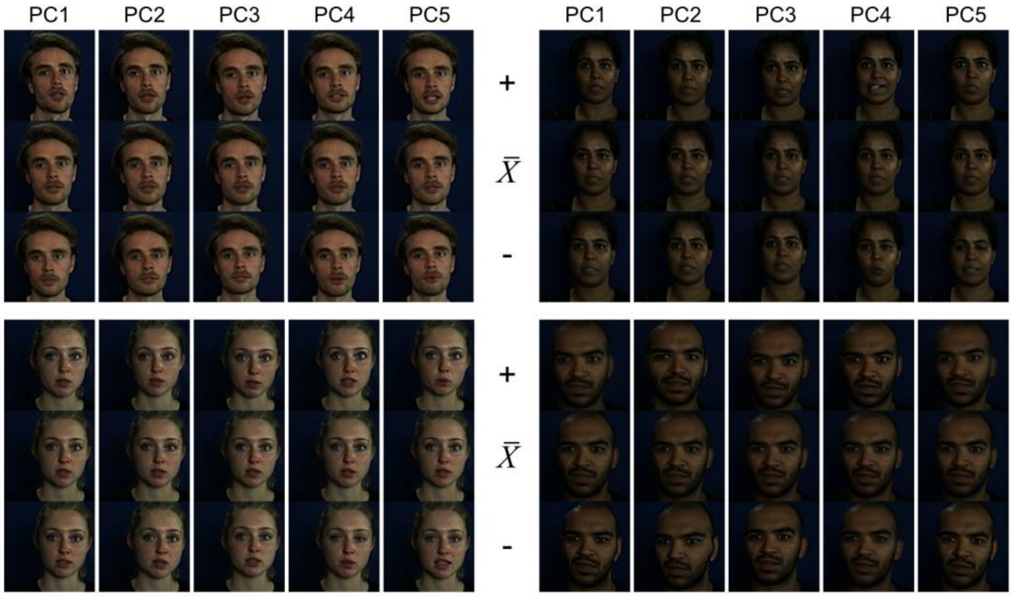


Figure 4.31. Example components across actors in Layer 2 of Model 5.

The first 5 PCs of the multi-view space of layer 2 for four actors, showing the front view only. In each of the 4 sets, the middle row represents the origin of the multi-view space. Top and bottom rows represents scaling the components by +3SD and -3SD respectively.

4.8.5.2.2 Reconstructions

As expected, reconstructions made from the combined space using the inner product loadings and no scaling were severely muted (Figure 4.32). They were no better than the unscaled reconstructions from Model 3 (loading similarity ED: Model 3: mean = 479.74, SD = 95.45; Model 5: 480.04, SD = 95.21; *t*(8) = 1.61, *p* = .147, 95% CI = [-0.13, 0.73]; loading similarity Z: Model 3: mean = 0.969, SD = 0.055; Model 5: 0.969, SD = 0.065; *t*(8) = 0.00, *p* = .997, 95% CI = [-0.015, 0.015]).

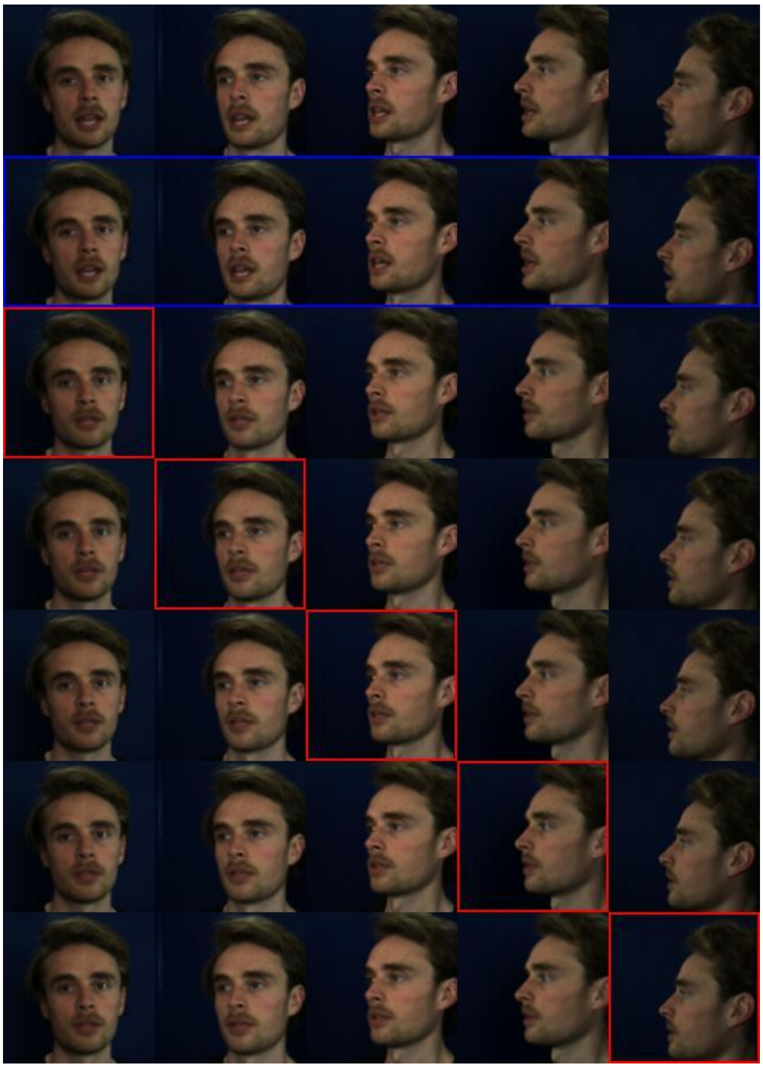


Figure 4.32. Example reconstructions from the second, view-independent layer of the two-step model, using inner products and no scaling.

Top row shows the veridical frames. Second row (bounded by the blue box) shows reconstructions made from the multi-view input. Each subsequent row shows the reconstructions made by projecting the view bounded by the red box into its respective space in layer 1, then projecting the reconstructed deviation vectors into the multi-view layer and reconstructing all other views. In these reconstructions, only the first 100 components of each space were used. As can be seen, the multi-view reconstructions are accurate, while single-view reconstructions are muted.

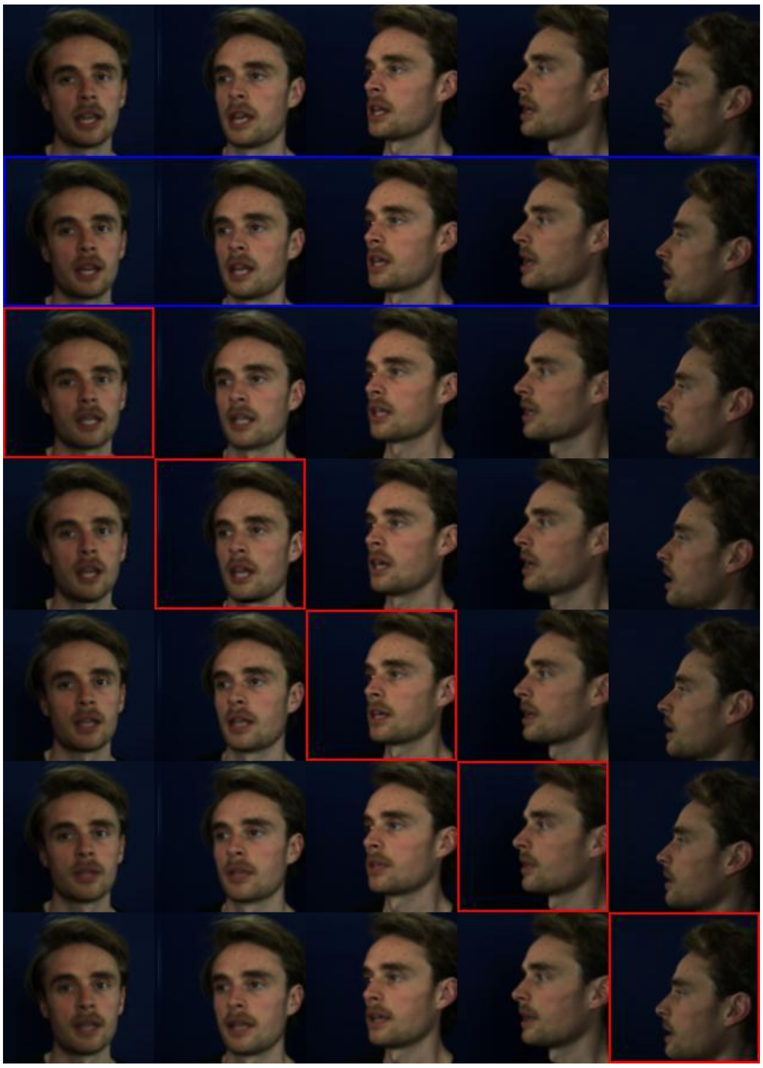


Figure 4.33. Example reconstructions from the second, view-independent layer of the two-step model, using inner products and simple scaling factors.

Top row shows the veridical frames. Second row (bounded by the blue box) shows the reconstructions made from projecting the multi-view vector into the PCA space. Each subsequent row shows the reconstructions made by projecting the view bounded by the red box into its respective space in layer 1, then projecting the reconstructed deviation vectors into the multi-view layer and reconstructing all other views. Only the first 100 components of each space were used, scaled by the simple scaling factors calculated using cascade reconstructions.

Reconstructions made using the simple scaling factors (λ), were substantially improved (see Figure 4.33). Movement was slightly muted across large changes in view but was sufficiently dynamic for reading speech. For trained frames, there was a significant main effect of scaling (no scaling, simple scaling (λ), the exponential fit of λ ($e\lambda$), and using relative rather than inner product loadings) on the loading similarity Z-value (*F*(1.41, 11.26) = 152.69, *p* < .001, η_p^2 = 0.95, Greenhouse-Geisser correction applied) and ED (*F*(1.74, 13.89) = 75.74, *p* < .001, η_p^2 = 0.90, Greenhouse-Geisser correction applied). Applying the exponential fit (Z: mean = 1.221, SD = 0.087, ED: mean = 338.97, SD = 68.66) significantly improved the reconstruction accuracy compared to no scaling (Z: mean = 0.969, SD = 0.065, *t*(8) = 11.94, *p* < .001, 95% CI = [0.203, 0.300], ED: mean = 480.04, SD = 95.21, *t*(8) = -8.61, *p* < .001, 95% CI = [-178.85, -103.30]). Applying the simple scaling factor (Z: mean = 1.278, SD = 0.093, ED: mean = 310.68, SD = 62.58) improved reconstruction accuracy compared to both no scaling (Z: *t*(8) = 12.25, *p* < .001, 95% CI = [0.251, 0.367], ED: *t*(8) = -9.81, *p* < .001, 95% CI = [-209.17, -129.55]) and exponential scaling (Z: *t*(8) = 4.82, *p* = .001, 95% CI = [0.030, 0.084], ED: *t*(8) = -4.12, *p* = .003, 95% CI = [-44.14, -12.44]). A comparison of scaling options is presented in Figure 4.34.

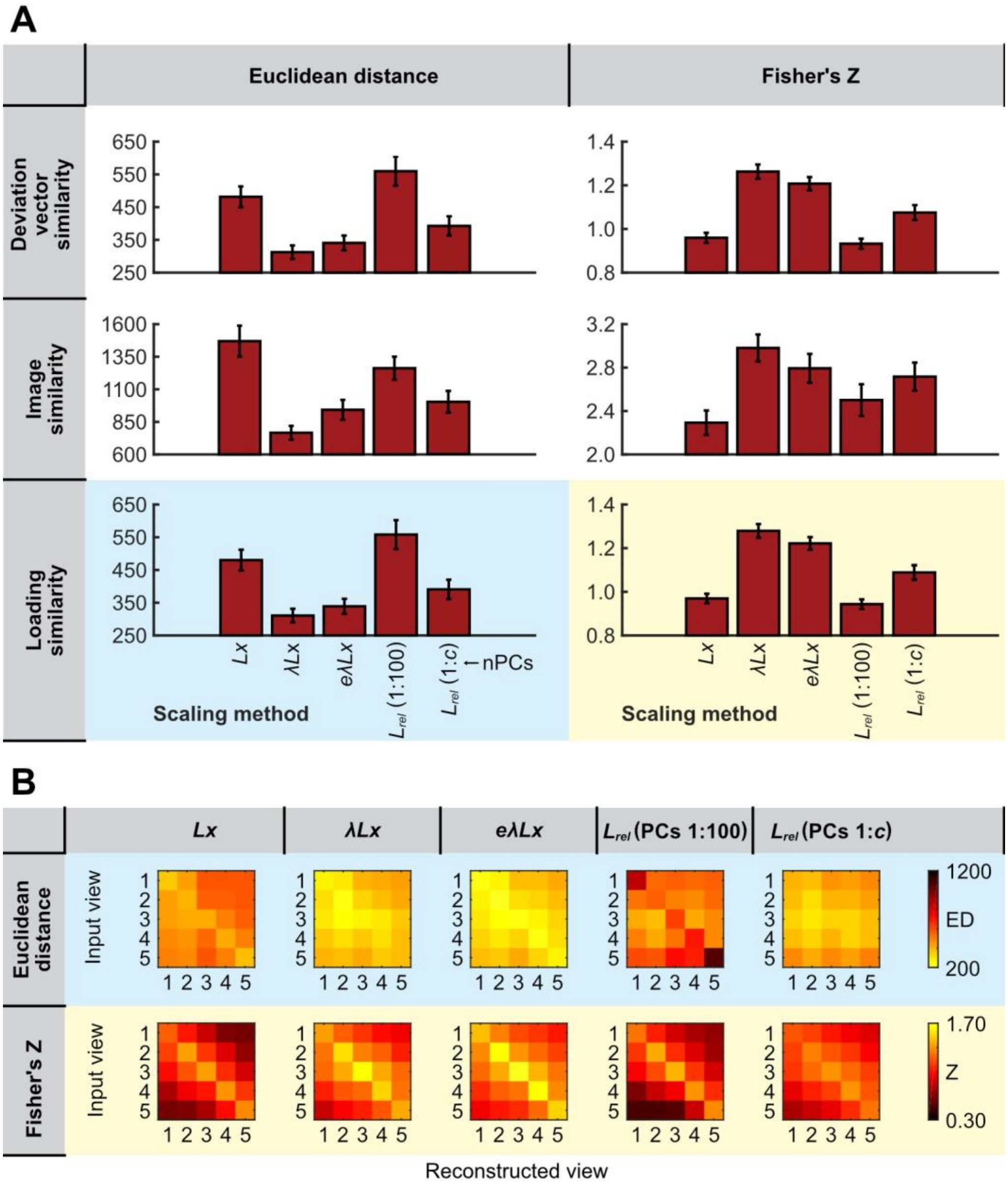


Figure caption on next page

Figure 4.34. Comparison of scaling methods in the two-step model.
(**A**) Mean reconstruction accuracy (ED – left, Z – right, deviation vector similarity – top, image-level similarity – middle, loading similarity - bottom) across actors for each scaling method. 'Lx' = no scaling. 'λLx' = applying simple scaling factors. '$e\lambda Lx$' = applying exponential scaling. 'L_{rel}' = using relative rather than inner product loadings with either all 100 PCs or the best c PCs. Error bars show ±1 SEM of the within-subject mean. (**B**) Loading similarity separated by view for each scaling method, showing the mean ED (top) and Z (bottom) across actors. In all heatmaps, brighter colours reflect better reconstruction accuracy.

What is surprising given the additional steps, but reassuring, is that the reconstructions were slightly but significantly more accurate on average than from Model 3 using $\beta 1 Lx + \beta 0$. For trained frames, the ED of the loading similarity was smaller (Model 5: mean = 310.67, SD = 62.58, Model 3: mean = 313.92, SD = 62.77, *t*(8) = -5.80, *p* < .001, 95% CI = [-4.54, -1.96]) and the Z was larger (Model 5: mean = 1.278, SD = 0.093, Model 3: mean = 1.265, SD = 0.094, *t*(8) = 6.49, *p* < .001, 95% CI = [0.008, 0.017]). These effects were also seen for untrained frames (ED: *t*(8) = -2.59, *p* = .032, 95% CI = [-10.39, -0.61]), Z: *t*(8) = 3.10, *p* = .015, 95% CI = [0.004, 0.026]), although neither result for untrained frames survived Bonferroni correction (α = .0125). Of course, the more complicated reconstructions in Model 5 may not always be better for different actors or videos captured under different circumstances. Nevertheless, it is reassuring that this model can make visibly accurate reconstructions as good as with the best scaling method in Model 3, despite additional steps to increase biological plausibility.

To further support the use of reconstructions from layer 1 for determining λLx scaling, we compared reconstruction accuracy to frames scaled using the veridical multi-view inputs. The accuracy was marginally worse when scaled using the cascade reconstructions, but the difference was negligible. The loading similarity Z-value was fractionally worse for cascade scaling (mean = 1.278, SD = 0.093) than veridical scaling (mean = 1.280, SD = 0.093, *t*(8) = -3.49, *p* = .008, 95% CI = [-0.003, -0.001]). The ED was not significantly different (veridical: mean = 310.41, SD = 62.42, cascade: 310.68, SD = 62.58, *t*(8) = 2.19, *p* = .060, 95% CI = [-0.01, 0.54]). Despite additional steps, the method was still effective.

The use of relative loadings rather than inner product loadings was also assessed, revealing the same problems as Model 3. Reconstructed motion was more visible than when using inner product loadings without scaling but was often caricatured as were the textures (Figure 4.35). Using relative loadings (Z: mean = 0.943, SD = 0.065, ED: mean = 558.02, SD = 132.08) rather than inner product loadings (Z: mean = 0.969, SD = 0.065, ED: mean = 480.04, SD = 95.21) with all 100 PCs resulted in worse reconstructions (Z: *t*(8) = -3.04, *p* = .016, 95% CI = [-0.046, -0.006]), ED: *t*(8) = 3.99, *p* = .004, 95% CI = [32.87, 123.09]).

The detrimental effect of using the relative loadings was again lessened if the spaces were cropped to the first c components (see Figure 4.34), with c determined separately for each actor to provide the best reconstruction accuracy within the first 20 components. The loading similarity (ED: mean = 390.70, SD = 88.89, Z: mean = 1.089, SD = 0.099) was improved over using all PCs (ED: *t*(8) -7.73, *p* < .001, 95% CI = [-217.22, -117.42], Z: *t*(8) 6.69, *p* < .001, 95% CI = [0.096, 0.196]) or using the inner product with no scaling (ED: *t*(8) -4.32, *p* = .003, 95% CI = [-137.08, -41.60], Z: *t*(8) 5.04, *p* = .001, 95% CI = [0.065, 0.175]). However, the reconstructions were still worse than using inner product loadings and simple scaling factors (ED: *t*(8) 6.94, *p* < .001, 95% CI = [53.45, 106.61], Z: *t*(8) -19.89, *p* < .001, 95% CI = [-0.211, -0.167]). Figure 4.36 shows an example using only the first 13 components of the multi-view space, showing the same frame (not used for training) as in figures Figure 4.29, Figure 4.32, Figure 4.33 and Figure 4.35. The correlation between the single-view loadings and the multi-view loadings also decreased with increasing component number. Using the relative loadings was therefore somewhat effective, but only for the first c components where the correlation between the single-view and multi-view loadings is highest.

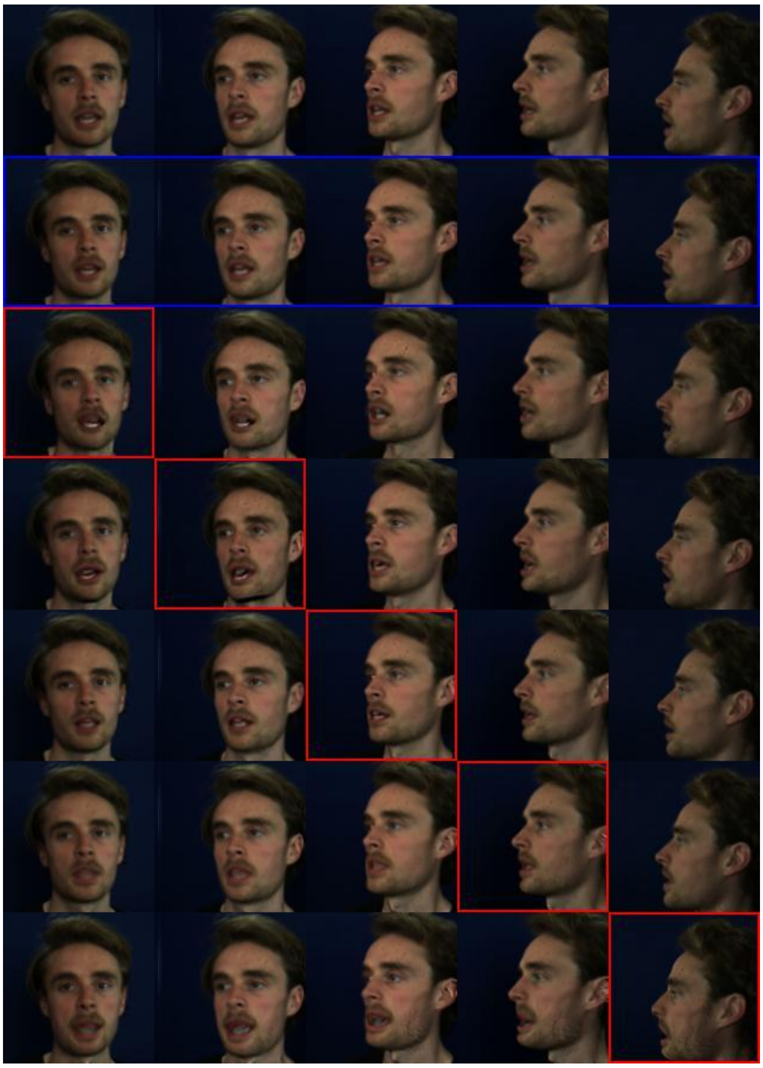


Figure 4.35. Reconstructions from the second, view-independent layer of the two-step model, using relative loadings for the first 100 components.

Top row shows the veridical frames. Second row (bounded by the blue box) shows the reconstructions made from the multi-view inputs. Each subsequent row shows the reconstructions made by projecting the view bounded by the red box into its respective space in layer 1, then projecting the reconstructed deviation vectors into the multi-view layer and reconstructing all other views. The multi-view reconstructions are accurate, while reconstructions from single-view inputs are slightly caricatured and, in some cases, distorted, e.g., see bottom row.

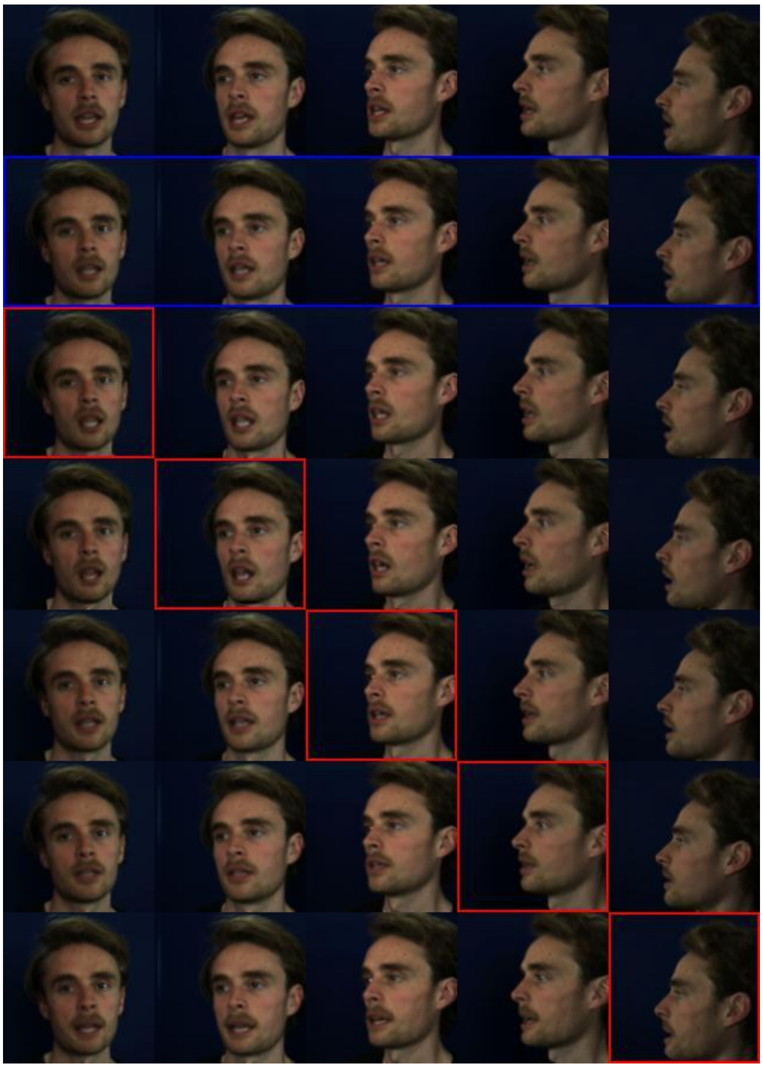


Figure 4.36. Example reconstructions from the second, view-independent layer of the two-step model, using relative loadings and only the first 13 components.

Top row shows the veridical frames. Second row (bounded by the blue box) shows the reconstructions made from the multi-view input. Each subsequent row shows the reconstructions made by projecting the view bounded by the red box into its respective space in layer 1, then projecting the reconstructed deviation vectors into the multi-view layer and reconstructing all other views. Reconstructions from single-view inputs are still slightly caricatured but less so when restricted to the first 13 components (Figure 4.35).

We next tested the generalisation to new frames. Again, reconstructions for untrained frames were visibly accurate but quantitatively worse. The loading similarity Z-values when scaled with simple scaling factors were not significantly different for untrained (mean = 1.225, SD = 0.168) than trained frames (mean = 1.278, SD = 0.093, *t*(8) = -0.76, *p* = .471, 95% CI = [-0.215, 0.109]). However, the ED was significantly worse (untrained: mean = 476.13, SD = 166.51, trained: mean = 310.68, SD = 62.58, *t*(8) = 3.73, *p* = .006, 95% CI = [63.12, 267.80]), suggesting a difference in magnitude rather than direction.

This pattern was also observed for reconstructions from the first layer of the model. The Z-value, ignoring same-view reconstructions, was not significantly different between trained (mean = 1.633, SD = 0.129) and untrained frames (mean = 1.448, SD = 0.233, *t*(8) = 1.69, *p* = .130, 95% CI = [-0.067, 0.437]). In contrast, the ED was significantly worse for untrained (mean = 416.13, SD = 181.76) than trained frames (mean = 220.90, SD = 45.69, *t*(8) = 3.29, *p* = .011, 95% CI = [58.31, 332.17]). As the angle was not significantly different this again suggests the difference was primarily in the magnitude.

4.8.6 Discussion

Here, the aim was to make a biologically plausible model that recapitulated the hierarchical transition from view-dependence to view-invariance (Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991). The first layer provided view-dependent representations while the second achieved view-invariance, with components containing information about all 5 views.

Despite the additional steps necessary to create the representation in a biologically plausible manner, the model was able to reconstruct motion well from either layer and could generalise to new frames reasonably well. It combined some of the techniques used by Lan and colleagues (2012) to learn the associations between separate spaces, and then formulated a multi-view representation as previously created by Beridze (2021).

If the aim were purely computational, one could ignore the second, multi-view layer. The reconstructions from the first layer were more accurate than all other models created here, including the second layer of this model. Thus, for transforming across views for automatic speech-reading software, for example, layer 1 of this model should be used.

The complete model comes with increased computational costs over the multi-view spaces of Model 3 or the separate spaces of Model 1. Firstly, there are increased storage costs of having both view-specific and multi-view representations, as well as storing information about how the separate spaces are related. In contrast, the multi-view space of Model 3 only has one set of ‘neurons’ (components) with the associations between different viewpoints already learned and implicitly stored within the representation. The two-step model also has additional processing demands, with two stages of representation and reconstruction rather than just one.

To reduce computational costs, one could assume the separate spaces are redundant and are pruned once the multi-view representations are learned. But how then would the multi-view model update with new information? It would also be inconsistent with the aims to imitate the hierarchical, posterior-anterior progression from view-dependence to view-invariance (Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991).

Overall, the two-step model was able to reconstruct motion well and could be constructed in a biologically plausible manner. The model neither needed to ‘see’ all views concurrently when creating the multi-view space or when determining scaling. Despite creating impressive cross-view reconstructions using the second layer of the model, the reconstructions were better using the first layer, therefore, the first layer of the two-step model is best if one’s goals are purely computational.

4.9 General discussion

This chapter aimed to expand on the work by Beridze (2021) to develop a biologically plausible model of view-invariance that can reconstruct facial motion across views during speech, as if performing a mental rotation. While the question on the importance of motion and 3D information in identification is still open (Hancock et al., 2000), we have shown that a view-invariant representation of facial motion is possible without the necessity to form a 3D model.

Model 1 comprised of separate spaces for each view and attempted to directly project deviation vectors from one view into the space of the neighbouring view. Model 2 tried to form a single space in which different views could be considered around local, view-specific means. Both were unsuccessful, with problems of spatial positioning and superimposition causing poor reconstructions.

Replicating the methods of Beridze (2021), Model 3 concatenated all 5 views into multi-view vectors to form one, multi-view PCA space which was able to reconstruct motion well across views once scaling was applied. While not biologically plausible, it provides a good computational model and provided a basis for making Models 4 and 5.

Model 4 attempted to build the multi-view space in a more biologically plausible way by only concatenating neighbouring views, with the hope that the overlap would allow associations to be extracted across all 5 views. It was able to learn some of the correlated movements across the different views. However, Model 4 had problems that could not easily be solved, with the motion in the components not always being consistent across views, with some reversals in the direction of actions and some different, non-reversed actions.

Model 5 took inspiration from Lan and colleagues (2012) and some of the previous models outlined by Ding and Tao (2016) using canonical correlation analysis, to learn the relationships between the components in the separate spaces for different views. The learned relationships, in the form of transformation matrices, were then used to transform the loadings and reconstruct frames across views. The reconstructions were then used to build a multi-view space and provide suitable scaling. Model 5 provided the most biologically plausible model, comprising 2 layers, that of all our models best mimicked the transition in neural tuning from view-dependence to view-independence (Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991). The first, view-dependent layer was able to reconstruct motion remarkably well, the best of all models created thus far. The second, view-independent layer was able to reconstruct motion across views but, like Model 3, the loadings needed scaling. Unlike Model 3, the scaling could be determined in a more biologically plausible way.

In Model 3 and the second layer of Model 5, the relative loading was also considered rather than the inner product loading. The aim was to use the magnitude of the truncated component to scale the loadings. The intuition was that this would prevent the need to project in either veridical or reconstructed multi-view vectors for determining scaling. However, when using all components, the reconstructions were often caricatured. Further investigation through assessing the correlations between the single view and multi-view loadings suggested this was due to the correspondence across the views decreasing as the component number increased, although it was not possible to see this discrepancy from visualising the components. Nevertheless, this indicates that the scaling factors identified also down-regulate the loadings to mediate the decreasing correspondence.

Using the relative loadings with only the early components, where the correlation between the single-view and multi-view loadings was higher, improved reconstruction accuracy compared to either using all components or compared to using the inner product loadings with no scaling, showing it was somewhat effective. However, the reconstructions were still not as good as using the scaling factors determined using either the veridical or reconstructed multi-view vectors.

Although not formally tested, caricaturing may not always be problematic. It was often easy to tell what the actor was saying when the motion was caricatured using an early subset of the components. Moreover, Furl and colleagues (2020) found that caricatured dynamic expressions were still recognisable and convincing, and exaggerating lip movements has also been found to improve human lip-reading (Theobald et al., 2006). Therefore, caricatured motion in speech may even be beneficial for some applications.

The first layer of the model adapted methods used previously (Lan et al., 2012; Lucey et al., 2007) to transform the appearance of the full face across views rather than just the mouth. It also improved the biological plausibility by not requiring the ‘observer’ to view all 5 views simultaneously unlike in Lan et al (2012) and Beridze (2021). The methods also have practical advantages. Because the associations between views are only learned between neighbouring views, it means only 2 rather than 5 cameras are required. Having more cameras makes the task easier as you do not need to record multiple videos, yet 2 will be sufficient with no need to record from every possible pair of views (e.g., 1-2, 1-3, 1-4, 1-5, 2-3, 2-4, 2-5 …).

The second layer of Model 5 also provides a plausible and accurate model for reconstructing facial motion across views. It has the same limitations as the multi-view model in Model 3 and previous work (Beridze, 2021), in that scaling is required to enhance the reconstructions. Unlike Model 3, however, there is no need to project all veridical views back into the space to ascertain the required scaling factors. Instead, the required scaling factors can be determined by projecting the multi-view reconstructions from the first layer into the space, retaining more biological plausibility.

The reconstructions were quantitatively worse for untrained than trained frames, yet they were still visibly accurate. Nevertheless, future work should look to improve the model so that the model can better generalise to new frames with smaller errors in the reconstructions. Here the videos were taken under highly controlled conditions. Training the models of videos captured under more variable conditions should improve the model’s capacity to generalise to new frames.

Despite saying the same sentences, the components in the models showed signs of being idiosyncratic. This is consistent with the work by Burton and colleagues (2016) yet poses a problem to the question outlined by Dobs and colleagues (2018) as to how many components are needed to encode the full space of facial motion. We found a less clear division between rigid and non-rigid motion compared to Burton and colleagues (2016) but this is unsurprising; our videos were captured under controlled conditions with actors repeating 10 sentences. Had the videos been taken under more varied conditions and with a variety of tasks we would likely have seen a stronger separation of rigid and non-rigid motion.

To summarise, we have created a 2-layer model that can reconstruct facial motion across views well whilst retaining biologically plausibility in not being exposed to multiple or distal views simultaneous. It also partially replicates the transition from view-dependence to view-independence seen in the macaque face processing system.

The above work focused on one side of the face, yet it is important to consider how the other side of the face, and mirror-views might be represented. This will be discussed in the next chapter.

4.10 Supplementary materials

4.10.1 Supplementary videos

Supplementary videos can be found using the link below:

https://www.dropbox.com/scl/fo/kaqognq18e0a5vira9syo/h?rlkey=m0k5mrhvglk69rpjdy92lk5ha&dl=0

Within the Supplementary Materials folder, navigate to 'Supplementary_Videos/Chapter_4_multiview_face_space'. This folder contains videos from all 5 models. Within each model there are folders containing 'reconstructions' and/or 'pc_videos' (videos of the principal components), sometimes for one actor only ('S1') or sometimes for two actors ('S1' & 'S3'). There are Word documents ('info.docx') provided throughout to provide information about the videos.

Chapter 5 A collapsed representation of motion across mirror views

5.1 Preface

In this chapter, the role of mirror views in face processing and some methods for incorporating mirror view representations into the multi-view PCA work are discussed. A pilot computational experiment is also presented demonstrating attempts at reconstructing mirror-flipped motion.

5.2 Introduction

The previous chapter presented models of view-invariant motion processing, culminating with a 2-layer model. The first layer contained separate PCA spaces for each viewpoint and the second a view-invariant, multi-view space. This work only included one hemi-view, so the next stage is to explore how mirror views might be incorporated into the model, and into face space more generally. For clarity, hemi-view refers to which side the face is seen from.

Previously, it was thought that neural responses to mirror views are a by-product of increased interhemispheric connectivity (e.g., see Corballis & Beale, 2020) yet more recently the emphasis has been on mirror views as a functional stepping stone to achieving view-invariance (Flack et al., 2019; Freiwald & Tsao, 2010; Meyers et al., 2015; Rogers & Andrews, 2022). Recent work using DNNs supports both suggestions (Farzmahdi et al., 2023) also emphasising the importance of spatial pooling across the visual field on the emergence of mirror responses.

This spatial pooling is consistent with interhemispheric connectivity leading to wider receptive fields (e.g., see Gross et al., 1977). However, while Farzmadhi and colleagues used a relatively simple DNN (AlexNet), a potentially simpler representation in face space might be possible. A simple active appearance model better predicted neural responses in macaque face patch AM than many DNNs, including AlexNet (Chang et al., 2021). The current chapter therefore explores how this intermediate stage for mirror views might be represented within our relatively simple multi-view face space.

The hierarchical progression to view-invariance has been shown in macaques (Freiwald & Tsao, 2010; Meyers et al., 2015) where the more posterior neurons in areas ML/MF show sensitivity to viewpoint, neurons in anterior area AM show invariance to viewpoint, and neurons in AL respond to mirror symmetric views.

The intermediate preference for mirror symmetric views has also been seen in humans. For instance, adaptation aftereffects transfer across mirror views (Jeffery et al., 2006, 2007) to a greater degree than can be explained by broad tuning curves, and response patterns in FFA and pSTS are more similar for mirror viewpoints than non-mirror views (Axelrod & Yovel, 2012; Flack et al., 2019). Similar neuroimaging results were also observed by Rogers and Andrews (2022), although, interestingly the right pSTS only showed an effect of mirror symmetry for familiar faces. The OFA, FFA and pSTS also showed more similar response patterns to symmetrical picture-plane rotations, suggesting that sensitivity to symmetry is a general property of the neurons. This is further supported by mirror responses emerging for both faces and other object categories in DNNs (Farzmahdi et al., 2023). This general property may help build viewpoint-invariance, but possibly in different ways depending on the symmetry. Responses to symmetrical views across changes in yaw *may* help build viewpoint-invariance through mechanisms discussed below, while responses to mirror symmetry in roll (picture plane) *may* help with alignment. This dissociation would of course need to be investigated.

While research has shown that neurons are sensitive to mirror symmetrical views, and researchers have suggested that mirror views are important for developing view-invariance, the mechanism by which this occurs is not yet clear. This chapter explores some possible mechanisms by which mirror views might be represented, particularly within the context of facial motion. Although not tested here, it is worth noting that the role and representation of hemi-views may differ for tasks such as speech processing compared to identity processing. Faces are generally quite symmetric in structure, yet speech and other facial actions are often quite asymmetric. Processing of motion and identity across mirror views might therefore require different mechanisms.

There are various theories of why facial motion is asymmetric, which will be touched upon briefly here. Notably, asymmetries in emission and perception differ based on the content, being either emotional or verbal. The right hemisphere model (H. S. Asthana & Mandal, 1997) posits that emotional expressions are presented more clearly and intensely on the left hemiface due to a right hemisphere dominance for processing and generating expressions. This results in a stronger driving force for emotive actions on the left side of the face. This was supported by evidence showing that the left hemiface moves more than the right during facial expressions (Nicholls et al., 2004). The valence model (see Nicholls et al., 2004) further predicts that the left hemiface is particularly biased for negative emotions. While Nicholls et al found no interaction between asymmetric movement and valence, participants rated the left side of the face (from the actor's perspective) to be sadder and the right happier (Nicholls et al., 2004). This effect was stronger when viewed veridically, compared to when mirror-flipped, suggesting a physical asymmetry in the motion. The presence of the perceptual bias even when mirror-flipped, however, also supports an asymmetry in perception, consistent with suggested hemispheric differences in emotion processing (Natale et al., 1983).

Similar theories have also been used to explain asymmetries in speech. During speech, the right side of the mouth (from the speaker’s perspective) opens more and for longer than the left (Graves et al., 1982; Jordan & Thomas, 2007; Nicholls & Searle, 2006). This is theorised to result from the left-hemisphere dominance for language (e.g., Graves et al., 1982). Word recognition from visual speech is also easier from the right hemiface than the left (Jordan & Thomas, 2007; Nicholls & Searle, 2006), consistent with greater right-sided motion.

Participants also, however, show better accuracy for visual speech presented in the right visual field, especially when that information contains the right hemiface (Jordan & Thomas, 2007), consistent with a left hemisphere advantage for processing speech.

The main question of interest here, however, is not *why* facial motion is asymmetric, but *how* (a)symmetries across mirror views are represented in the context of facial motion? The use of mirror views for view-invariance might depend on whether faces are represented in a 2D or 3D manner. If the brain holds a 3D representation, one route to view-invariance might be through inverting the 3D structure of a seen view through the midsagittal plane to estimate the structure of the other hemiface and therefore the full structure of the face. But the additional complexity of determining 3D structure means this process is unlikely used for rapid, dynamic changes in facial motion. Instead, we focus on 2D representations which are likely faster, and explore how mirror views might be incorporated into the multi-view face space model.

The first consideration is that mirror views are not processed in the same way for motion as identity. Facial motion may instead be processed in a purely view-dependent manner. Research showing responses to mirror views typically uses static faces (Chang & Tsao, 2017; Flack et al., 2019; Freiwald & Tsao, 2010; Meyers et al., 2015; Rogers & Andrews, 2022) so it is currently unknown if mirror-responses emerge for dynamic stimuli. The pSTS, however, which is thought to be involved in processing dynamic facial aspects such as speech (Haxby et al., 2000; Pitcher & Ungerleider, 2021), is sensitive to mirror symmetry (Axelrod & Yovel, 2012; Flack et al., 2019; Rogers & Andrews, 2022), suggesting there is a role of symmetry in processing facial motion. This could be a biproduct of larger receptive fields and increased interhemispheric connectivity (e.g., see Corballis & Beale, 2020; Gross et al., 1977; Pitcher et al., 2020) given that unlike the OFA and FFA, the pSTS is equally sensitive to both ipsilateral and contralateral visual fields (Nikel et al., 2022; Pitcher et al., 2020). However, the sensitivity to mirror views in the pSTS is also modulated by familiarity (Rogers & Andrews, 2022) suggesting that familiarity might help build mirror-tuned representations.

Mirror views may help create templates during learning. There is sufficient evidence that the brain first represents each hemi-view separately (e.g., Freiwald & Tsao, 2010; Meyers et al., 2015), thus, during learning, templates for the unseen side might be estimated, through exploiting mirror symmetry to construct ‘virtual views’ (Vetter et al., 1994). This, however, is likely not the sole use of mirror views, as mirror responses to familiar faces suggest the representation is not restricted to learning (Rogers & Andrews, 2022). Moreover, the intermediate stage of mirror responses in more anterior areas (Meyers et al., 2015) suggests they have a bigger role to play in achieving view-invariance.

It may not be the case that the brain collapses across mirror views *per se*, but responses to mirror views might allow better access to a view-invariant representation. In this case, as depicted in Figure 5.1, the multi-view PCA space would have separate view-specific spaces for n viewpoints between ±90°, and the multi-view space would have separate ‘slots’ for them all. Mirror responses may be observed because inputs are then projected into both the veridical and mirror spaces, consistent with pooling across a larger receptive field and across hemispheres (e.g., see Corballis & Beale, 2020; Gross et al., 1977). Subsequently the representations from both veridical and mirror spaces might be projected onto the corresponding slots of the multi-view space, allowing better access to the view-invariant representation.

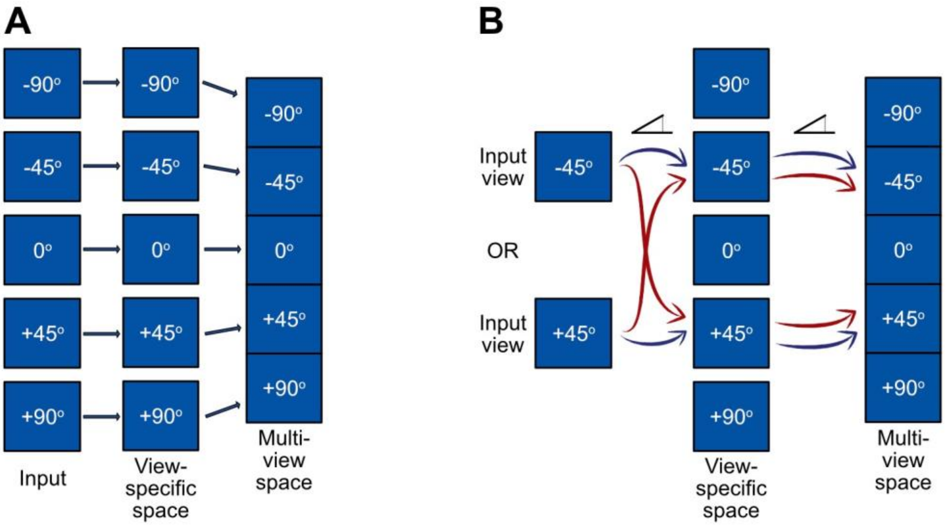


Figure 5.1. An example of a separate space model of mirror view processing.

(**A**) A separate PCA space is first made for each view. The multi-view space is then constructed with dedicated slots for both hemi-views. (**B**) Input views can be projected into both veridical and mirror-flip spaces, and the output from both spaces can be projected into the multi-view space.

Alternatively, facial motion may only be represented from ‘one side’, reducing computational and storage costs compared to representing both. In the context of the two-step multi-view model, the first layer would only contain separate spaces from 0^o to 90^o. One option for this is to only learn motion from one hemi-view, e.g., encoding the left hemi-view and ignoring the right. This of course is unlikely given that faces are generally viewed from both sides, but it is worth exploring. For this model, and the model with separate spaces for each hemi-view, the view-specific representations would need to code motion sufficiently well from one hemi-view, to be able to effectively process and recognise speech from both. This is especially important if motion is only learned from one side.

We therefore explored how well facial motion could be reconstructed in its mirror view’s space. We created separate spaces for frames in their veridical and mirror viewpoints and then projected the frames into the opposite spaces. If symmetrical enough, then the facial motion should be recoverable regardless of whether the mirror-flip or the veridical view is projected into a given space. If asymmetrical, then it will be challenging to recover the mirror-flipped motion as the spaces can only recover textures and motion on the span of the training set. As an extreme example, if using view-specific spaces one cannot recover a frontal image solely from a profile view’s space; the profile space simply does not contain the appropriate information to represent the frontal view. Due to the asymmetries in facial motion, we hypothesised that reconstruction fidelity would be better in the corresponding view’s space than in the opposite view’s space. We expected that symmetric motion would be reconstructed well from the opposite space, but asymmetric motion would not be. This example is outlined under the ‘separate spaces’ section of this chapter.

In the model with separate spaces for each hemi-view, it is difficult to imagine how the view-specific spaces could be learnt from one hemi-view and not update over time to also represent the other if both are subsequently being projected into the space. If they evolve to represent both hemi-views, then it would result in two spaces coding virtually the same information, the only difference being the direction the face is pointing. A more compact alternative with fewer redundancies would be to collapse mirror views into one representation with a separate signal coding left-right direction.

In the next experimental model, we therefore explored the possibility of a collapsed representation. In this case, the separate spaces from 0^o to 90^o would each contain information about both hemi-views (see Figure 5.2). As not all motion is asymmetric, it may be computationally advantageous to collapse across mirror views in this way. Symmetric motion such as a simple vertical drop of the mouth can be represented by one component, reducing computational costs compared to representing that same motion in separate spaces for left and right-facing views. Asymmetric motion might be represented by different components that are more hemi-view specific, or by opposing directions along single components.

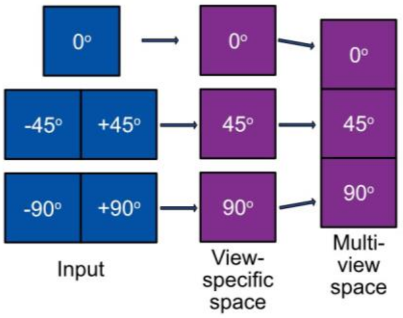


Figure 5.2. An example of a combined space model of mirror view processing. Mirror views are concatenated and combined. A separate PCA space is made for each combined view. The multi-view space is then constructed.

To simulate this, we created a single PCA space using frames from both the veridical and mirror-flipped view. We hypothesised that there would be separate components reflecting symmetric motion, structural asymmetries, and asymmetric motion. We then projected frames containing either the veridical or mirror-flipped motion into the space and assessed the reconstruction fidelity, hypothesising that we would be able to reconstruct the motion for both well.

5.3 Methods

5.3.1 Videos

New videos were acquired for this pilot experiment as the videos from the multi-view PCA work were illuminated and captured from one side only. In the current experiment, the illumination needed to be equal on both sides of the face. New videos were recorded of one actor (the author) with illumination provided through two diffuse lamps (InterFit F5, 50 x 69 cm), positioned at -25° and +25° yaw from frontal, with the centre of the lamps 20° above the camera position, approximately 128cm from the actor. In this pilot experiment, the actor was recorded from frontal (0°), -25°, and +25° while repeating the sentences from the multi-view PCA work.

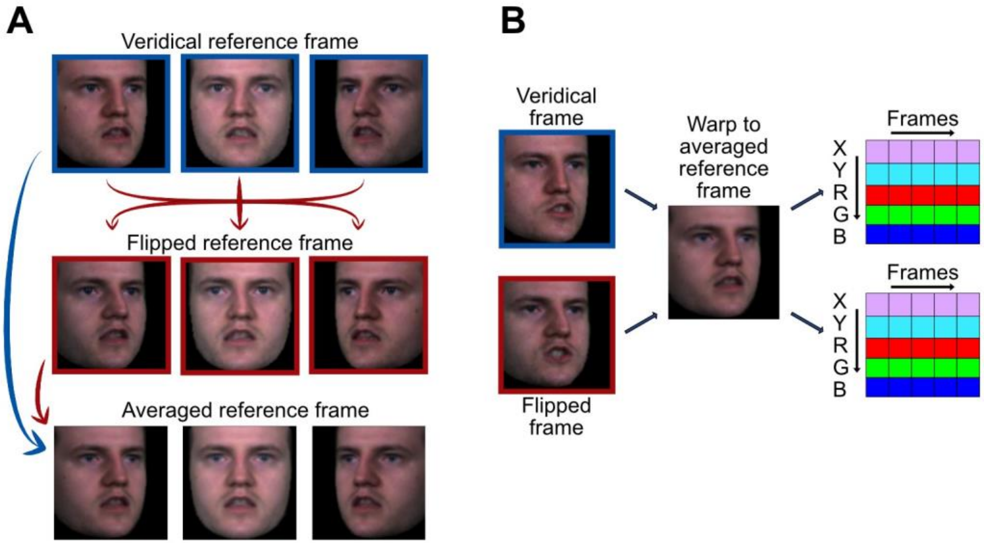


Figure 5.3. Mirror-flipping and warping to an averaged reference frame.

(**A**) The veridical (blue) and the mirror-flipped (red) reference frame for each view. The bottom row shows the averaged reference frame for each view, calculated in image space. (**B**) Warping the veridical and flipped frames to the averaged reference frame for the right-facing view. The columns of the 'XYRGB' matrices represent individual frames. The X (lilac) and Y (cyan) squares represent the X and Y warp fields from the McGM analysis. The R, G, and B squares represent the red, green and blue texture maps once aligned to the reference frame. Each individual square represents a 12,544 x 1 column vector (vectorised from the 112 x 112 pixels in image space).

The frames (N = 4450) were first spatially aligned using OpenFace (Baltrusaitis et al., 2018), so that when mirror-flipped, the faces would occupy the same regions of the image. Once aligned and down-sampled to 112 x 112 pixels, the frames were mirror-flipped, providing three pairs of frames: veridical and mirror-flipped frontal frames, veridical and mirror-flipped left-facing frames (mirror-flipped originally right-facing) and veridical and mirror-flipped right-facing frames (mirror-flipped originally left-facing). See Figure 5.3.

For each view (frontal, left-facing, right-facing), both the veridical and mirror-flipped frames were aligned to an averaged reference frame (Figure 5.3) using the McGM (Johnston et al., 1992, 1999). The template frame was an average of a veridical reference frame and its mirror-flip counterpart. The frame was chosen to have a relatively neutral expression with the mouth open, and the head positioned centrally to provide sufficient overlap between the veridical and mirror-flipped frames for each view. The veridical and mirror-flipped reference frames were averaged in image space. The same frame number was used for each view.

An averaged template was used for two reasons. Firstly, the ‘combined space’ model required all frames (veridical and flipped) to be aligned to a single template. Aligning to either the veridical or flipped template would bias the reconstructions towards that view. Secondly, in the ‘separate space’ model, it improved the alignment when projecting mirror-flipped frames into the veridical space and vice versa, minimising the impact of certain asymmetric features, such as the actor’s crooked nose. As demonstrated later in Figure 5.7., the asymmetrical nose is reconstructed well.

5.3.2 Separate space model

Separate PCA spaces were made for veridical and mirror-flipped frames. The X/Y fields and RGB textures were vectorised and concatenated as in the multi-view PCA models, such that each frame was a 62,720 x 1 column vector. For each view, the average vectors for the veridical ($\bar{V}$) and flipped ($\bar{F}$) frames were then subtracted providing zero-centred deviation vectors. PCA was performed on the veridical and flipped deviation vectors separately. All spaces were cropped to the first 100 components. Both veridical and mirror-flipped frames were then projected into the spaces.

To compare how well motion could be reconstructed in the matching or opposite spaces, we compared three reconstruction methods (see Figure 5.4). The first reconstruction method was veridical frames reconstructed from the veridical space. For this, $\bar{V}$ was subtracted from the veridical frames giving deviation vectors, which were projected into the veridical space. To reconstruct the frames, $\bar{V}$ was added to the weighted sum of the components and the image de-warped. The second method was the veridical frames (-$\bar{V}$) reconstructed in the space for the flipped frames. The reconstructed deviation vectors were a weighted sum of components in the flipped space. To reconstruct the frames $\bar{V}$ was added.

The third was the flipped frames projected into the veridical space. However, these were more like ‘pseudo-reconstructions;’ the aim was not to reconstruct the flipped frames *per se*, but to animate the veridical view with the flipped motion. If symmetrical, the reconstructed motion should be similar when animated with either the veridical or mirror-flipped view. For this, $\bar{F}$ was subtracted from the flipped frames and the deviation vectors projected into the veridical space. To construct the images $\bar{V}$ was added to animate the veridical view.

We hypothesised that reconstructions of veridical frames would be worse when made from the flipped space than the veridical space. We further predicted that the pseudo-reconstructions of the flipped frames would be less similar to the veridical frames than the veridical reconstructions in either space. To reiterate, if motion is perfectly symmetrical then the non-rigid motion of the mirror-flipped and veridical views should be identical and therefore recoverable in either space. If asymmetrical, then frames animated with the mirror-flipped motion should be less like the veridical frames, and the motion of veridical frames will only be partially recoverable from the flipped space.

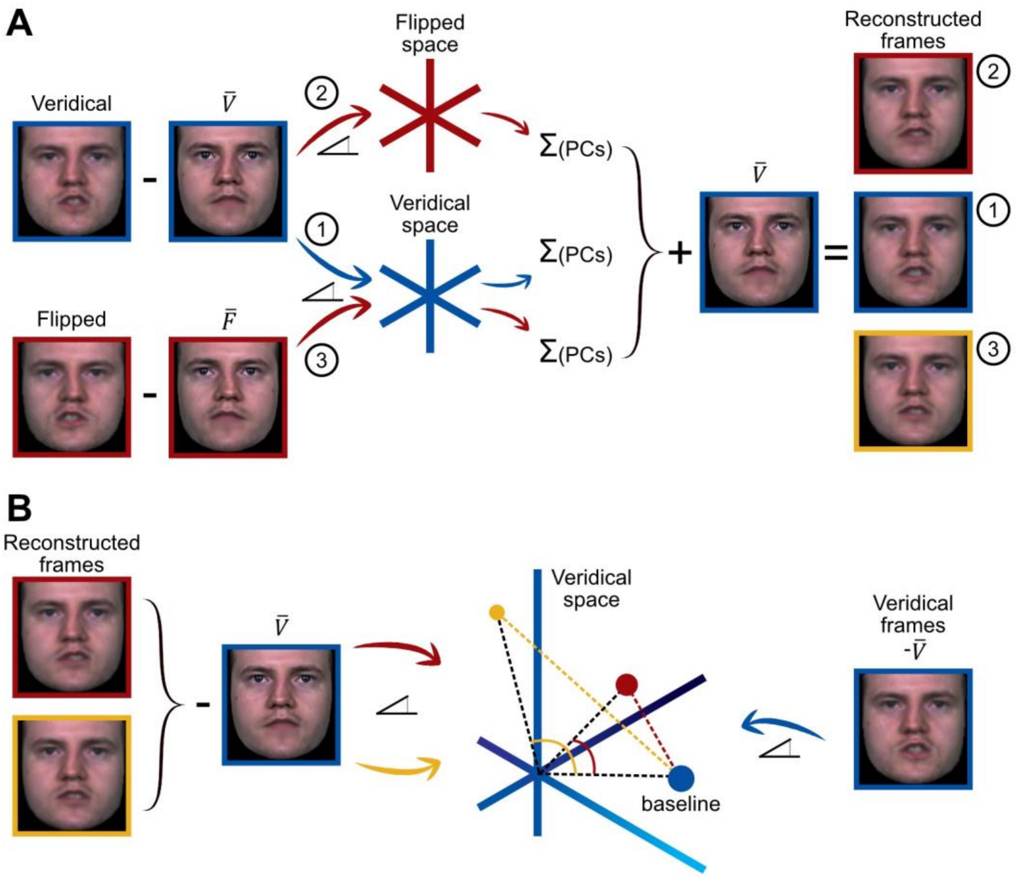


Figure 5.4. Projection and reconstruction in the separate spaces for mirror views.

(**A**) Veridical (blue) and mirror-flipped (red) frames are projected into the veridical and mirror-flipped spaces. The weighted components are summed to reconstruct deviation vectors. The average for the veridical frames is then added to reconstruct: the veridical frames projected into the veridical space (1, blue), the veridical frames projected into the flipped space (2, red), and the flipped frames projected into the veridical space (3, yellow). (**B**) Reconstructions are projected into the veridical space and compared to the baseline (the projection of the veridical frames into the space) to compare loading similarity. Images show actual frames and reconstructions. Projections into the veridical space in (**B**) illustrate the methods and predicted results, not actual data. The reconstructions are colour coded to match with the results.

5.3.3 Combined space model

In the ‘combined space’ model the X/Y fields and RGB textures for the veridical and mirror-flipped frames were vectorised and concatenated in time, akin to how different viewpoints were concatenated in Model 2 of the multi-view face space work. The average ($\bar{C}$) across the full set (N = 8900) was subtracted and PCA was performed. This was done for each view separately providing one space for frontal frames, one for left-facing, and one for right-facing frames. All spaces were cropped to the first 100 components. Veridical and mirror-flipped frames were projected into the spaces to assess reconstruction fidelity. We also inspected the components to assess how (a)symmetries were encoded.

Firstly, we assessed if the reconstructions of veridical frames from the combined space were as good as from the veridical space and if they were better than from the flipped space. To test this, we compared four reconstruction methods (see Figure 5.5). The first two reconstruction types were the reconstructions of veridical frames from 1) the veridical space and 2) the flipped space, as described in the ‘separate space’ model.

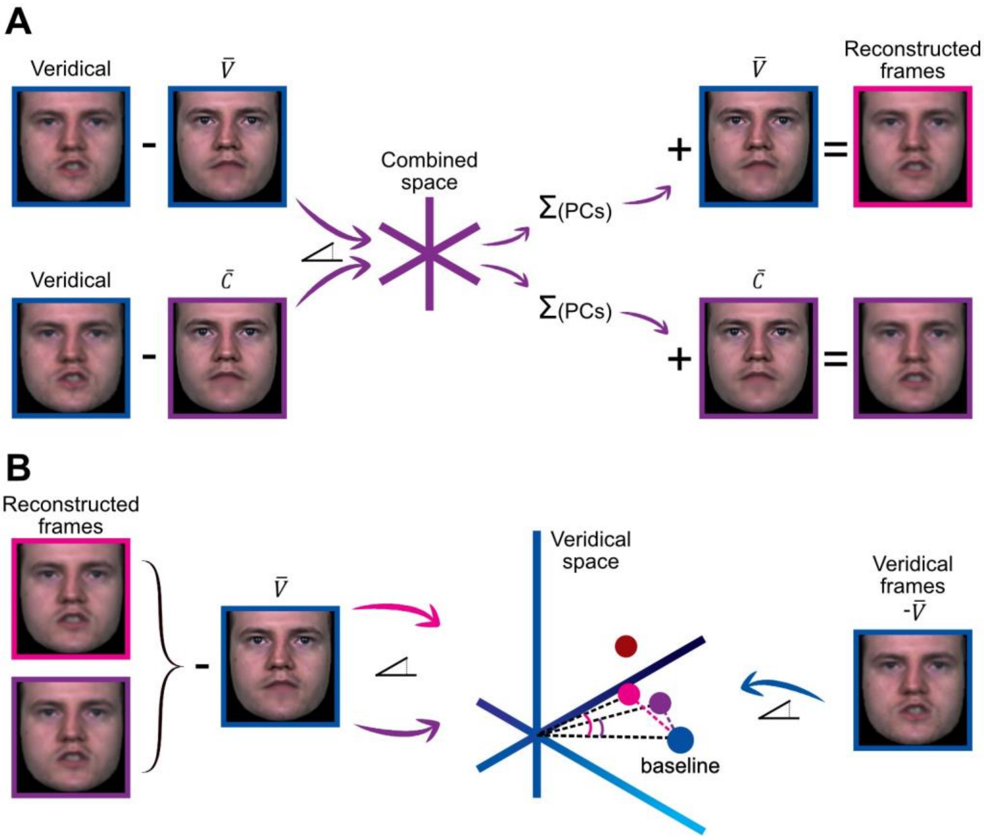


Figure 5.5. Projection and reconstruction in the combined space for mirror views.

(**A**) Veridical frames were projected into the combined space having either subtracted the average of the veridical frames (veridical mu) or the average across the veridical and mirror-flipped frames (combined mu). The weighted components were then summed to reconstruct deviation vectors. The veridical and combined averages were added to reconstruct the frames. (**B**) Reconstructed frames were projected into the veridical space and compared to baseline (the projection of the veridical frames into the space). The red dot shows the projection of the veridical frames reconstructed in the flipped space for comparison. Images show actual frames and reconstructions. Projections into the veridical space in (**B**) illustrate the methods and predicted results, not actual data. The reconstructions are colour coded to match with the results.

In the other two reconstruction methods, frames were projected into the combined space and varied subtly by whether $\bar{V}$ (3) or $\bar{C}$ (4) was subtracted prior to and re-added after projection. The difference being that deviation vectors for the veridical frames when $\bar{V}$ has been subtracted are devoid of any information about structural asymmetries. In contrast, $\bar{C}$ contains half-strength, averaged information about the view (veridical or flipped), thus subtracting $\bar{C}$ induces asymmetries into the deviation vectors. In either case, reconstructions should be accurate.

5.3.4 Measures of reconstruction accuracy

To quantify reconstruction fidelity we used the same measures of reconstruction accuracy as in the multi-view PCA work. We used the image and deviation vector similarity between the reconstructions and the original frames. We also compared loading similarities within the veridical spaces, by projecting the reconstructions (-$\bar{V}$) into the veridical spaces. For most comparisons, the loading similarity was measured relative to baseline (the projection of the veridical frames into the space). For each of these features, we assessed the Euclidean distance (ED) and the correlation coefficient (Fisher's Z) between either the pixels or the loadings.

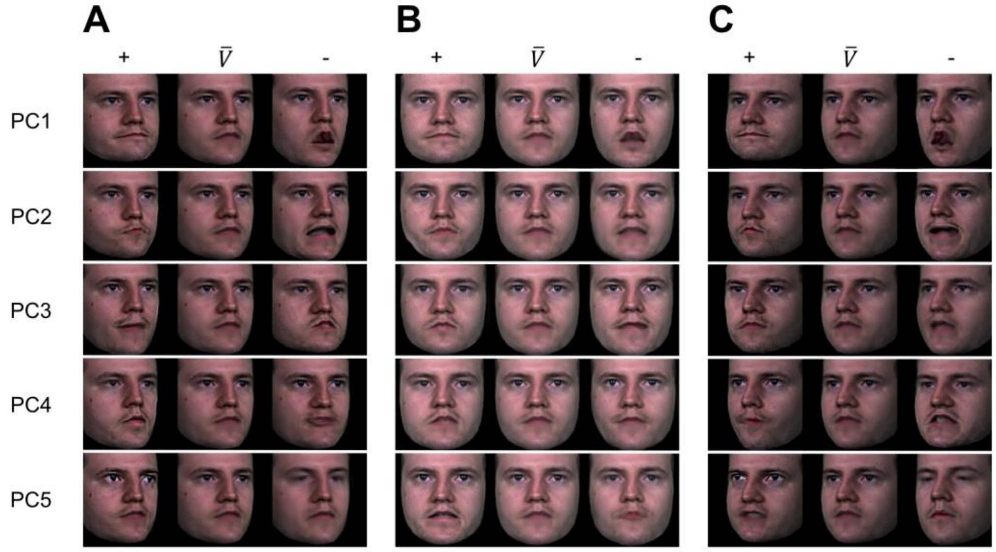


Figure 5.6. Principal components in the separate spaces for mirror views.

PCs 1 to 5 for right-facing (**A**), frontal (**B**) and left-facing (**C**) views. In each tuple, the middle image is the origin of the space. The left and right images show the polarities of the given component, created by scaling the unit vector for the component by +1000 and -1000 respectively. Note that while some components are similar, there are some differences across the three views.

5.4 Results

5.4.1 Separate spaces

5.4.1.1 Principal components

Example components for the three views can be seen in Figure 5.6. While some components are very similar across the three views, such as PC1, others differ, for instance PC4 which is different across all three viewpoints.

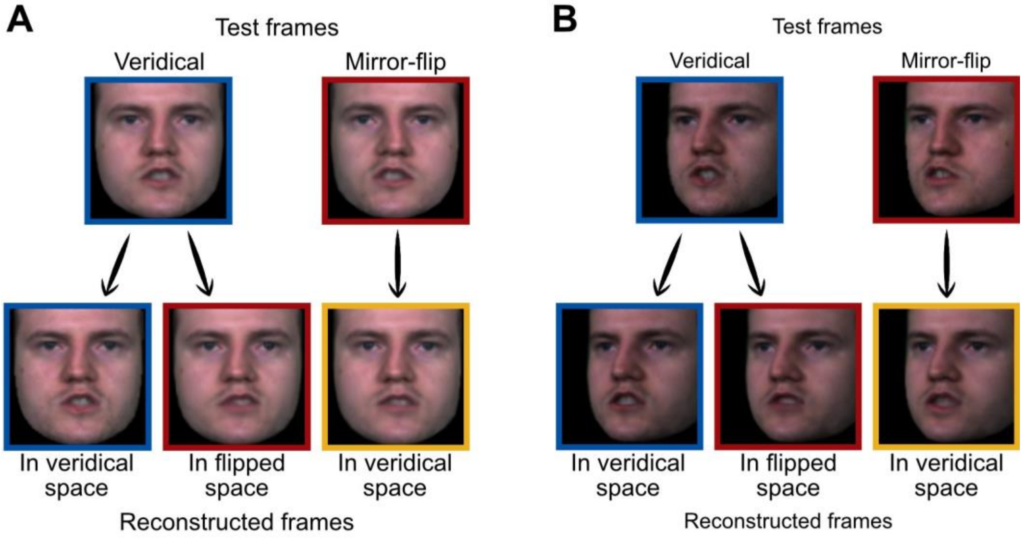


Figure 5.7. Reconstructed frame from the separate space model of mirror view processing.

Reconstructions of veridical and mirror-flipped frames in the veridical and mirror-flipped spaces. The top row shows the veridical (blue) and mirror-flipped (red) frames for frontal (**A**) and left-facing (**B**) views. The bottom row shows the reconstructions of the veridical frames in the veridical (blue) and flipped (red) spaces, and the pseudo-reconstructions of the mirror-flipped frame in the veridical space (yellow). The coloured borders correspond to the colours used in the projection and reconstruction methods in Figure 5.4 and the results in Figure 5.8.

5.4.1.2 Reconstructions

We were able to reconstruct the veridical motion well from the veridical space (e.g., see Figure 5.7). In contrast, reconstructions of either the veridical frames from the mirror-flipped space or the mirrored-frames from the veridical space were not as good. Note how the asymmetric lip separation is captured well using the veridical space, but less well using the mirror-flipped space. This is likely because the actor’s mouth consistently opens/closes asymmetrically. Some actions may be presented asymmetrically at a given time but alternate in laterality, and these should be reconstructed well even across mirror views. In contrast, other asymmetric actions such as the opening and protrusion of the lips in the example are more consistently unilateral, with the actor's mouth seldom opening more on the right side than the left. As a result, we cannot accurately recreate unilateral asymmetric actions from the flipped space.

Oddly, the asymmetric lip-separation was obvious in the reconstruction of the mirror-flipped frame in the veridical space for left-facing stimuli, despite not being a good reconstruction of the actual mirror-flipped frame. This incidental reconstruction of the veridical appearance from the mirror-flipped frame, however, cannot have been a regular occurrence, as (as will be detailed shortly) these pseudo-reconstructions were more dissimilar to the veridical frames than the reconstructions from the flipped space.

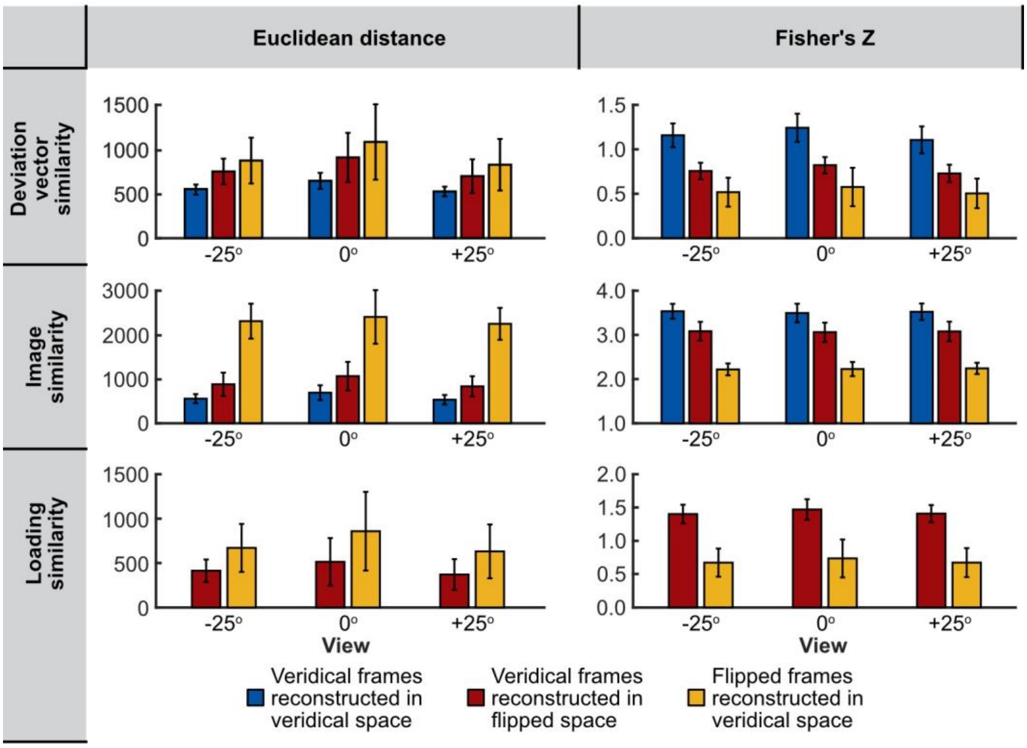


Figure 5.8. Reconstruction accuracy of veridical and mirror-flipped frames in the separate spaces.

Reconstruction accuracy (ED and Fisher's Z) for veridical frames reconstructed in the veridical (blue) and mirror-flipped spaces (red) and mirror-flipped frames reconstructed in the veridical space (yellow). Deviation vector and frame similarity was measured relative to the veridical frames. Loading similarity was measured relative to the loadings of the veridical frames in the veridical space. The colour code corresponds to the projection and reconstruction methods in Figure 5.4 and reconstructions in Figure 5.7. Error bars show ±1SD.

To quantify whether reconstructions were worse when projected into the opposite space, for each feature (image and deviation vector similarity) and measure (ED and Z) we performed a one-way within-subjects ANOVA using the frames as the 'subjects', and 3 levels of reconstruction method. The first and second being the reconstructions of veridical frames from the veridical and mirror-flipped spaces respectively. The third was the mirror-flipped frames reconstructed from the veridical space. Reconstruction accuracy is presented in Figure 5.8. The data was often skewed towards 0 for ED and was sometimes skewed in either direction for Z. ANOVAs and t-tests are robust to normality violations (Knief & Forstmeier, 2021; Schmider et al., 2010), however, as a precautionary measure Wilcoxon Signed Rank tests were performed alongside the post-hoc t-tests to confirm the findings.

We found that for frame and deviation vector similarity there was a main effect of reconstruction method for all views and all measures (all $F(2, 8898) > 5800$, $p < .001$, $\eta_p^2 > .57$). In all cases, the ED was lower for veridical frames reconstructed in the veridical PCA space compared to the flipped space (all $p < .001$) and the Z-value was higher (all $p < .001$) therefore providing better reconstructions. As an example, the ED of the image similarity for the frontal frames was smaller for the veridical frames reconstructed in the veridical space (mean = 688.08, SD = 167.34) than those reconstructed from the flipped space (mean = 1062.87, SD = 321.59, $t(4449) = -118.77$, $p < .001$, 95% CI = [-380.98, -368.60]). The Z-value of the image similarity for the same frames was higher for reconstructions from the veridical space (mean = 3.489, SD = 0.210) than the flipped space (mean = 3.058, SD = 0.218, $t(4449) = 220.84$, $p < .001$, 95% CI = [0.427, 0.435]).

The pseudo-reconstructions of the flipped frames in the veridical space were more dissimilar to the veridical frames than the veridical reconstructions from either space (all $p < .001$). For instance, the image similarity for the frontal frames was lower for the pseudo-reconstructions (ED: mean = 2400.08, SD = 605.61, Z: mean = 2.225, SD = 0.159) than the corresponding veridical frames reconstructed in either the veridical space (ED: *t*(4449) = 213.33, *p* < .001, 95% CI = [1696.27, 1727.74], Z: *t*(4449) = -439.73, *p* < .001, 95% CI = [-1.270, -1.258]) or the flipped space (ED: *t*(4449) = 208.52, *p* < .001, 95% CI = [1324.64, 1349.79], Z: *t*(4449) = -311.19, *p* < .001, 95% CI = [-0.838, -0.828]).

We also compared the loading similarity when the reconstructions were projected into the veridical space relative to baseline (the loadings of the veridical frames) between 1) the reconstructions of veridical frames from the flipped space, and 2) the pseudo-reconstructions of the flipped frames. The loadings of the veridical frames reconstructed from the flipped space (ED: mean = 515.43, SD = 267.45, Z: mean = 1.473, SD = 0.154) were more similar to the baseline than the loadings of the pseudo-reconstructions of the flipped frames (ED: mean = 860.68, SD = 443.82, *t*(4449) = -98.18, *p* < .001, 95% CI = [-352.15, -338.36], Z: mean = 0.738, SD = 0.285, *t*(4449) = 182.21, *p* < .001, 95% CI = [0.728, 0.743]).

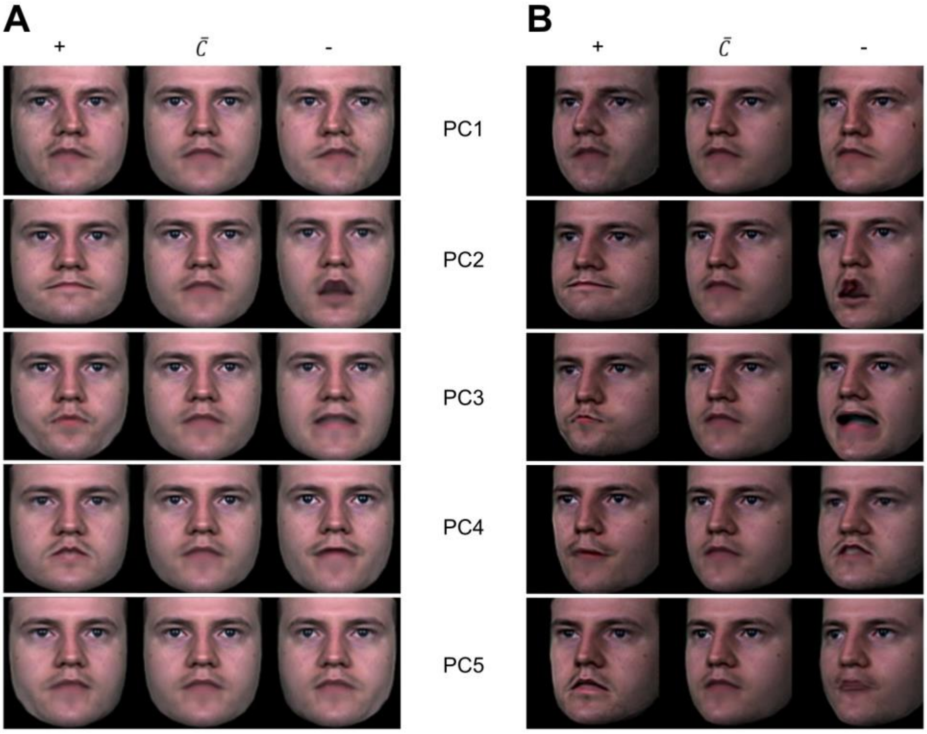


Figure 5.9. Principal components in the combined mirror space. PCs 1-5 in the frontal (**A**) and non-frontal (**B**) spaces. In each tuple, the middle image shows the origin of the space, and the left and right images show the components by +1000 and -1000 respectively. Note how PC1 seems to carry most of the information about hemi-view, while all other components are more symmetrical, with a straight nose and a mole of both cheeks. For reference, the mole is on the actor's right cheek.

5.4.2 Combined space

5.4.2.1 Principal components

We first inspected the components to assess the encoding of asymmetric and symmetric information. As expected, the origin is an averaged, symmetrical face. As presented in Figure 5.9, PC1 for both the frontal (0^{o}) and non-frontal (22.5^{o}) spaces encodes the structural asymmetries of the face almost fully. One side of the component reflects the left hemi-view and the other the right. Subsequent PCs coded symmetric and asymmetric non-rigid deformations.

5.4.2.2 Reconstructions

We then assessed reconstruction fidelity. As shown in Figure 5.10, we can reconstruct both dynamic and structural asymmetries reasonably well. ANOVAs were performed to test if reconstructions from the combined space were as good as or better than those made in the veridical and flipped spaces respectively. For each feature (frame similarity, deviation vector similarity), measure (ED, Z) and view, we performed a one-way within-subjects ANOVA with 4 levels of the independent variable, being the reconstruction methods. The first two were the reconstructions using the veridical (1) and mirror-flipped space (2). The other two were reconstructions from the combined space, created by subtracting and adding either $\bar{V}$ (3) or $\bar{C}$ (4).

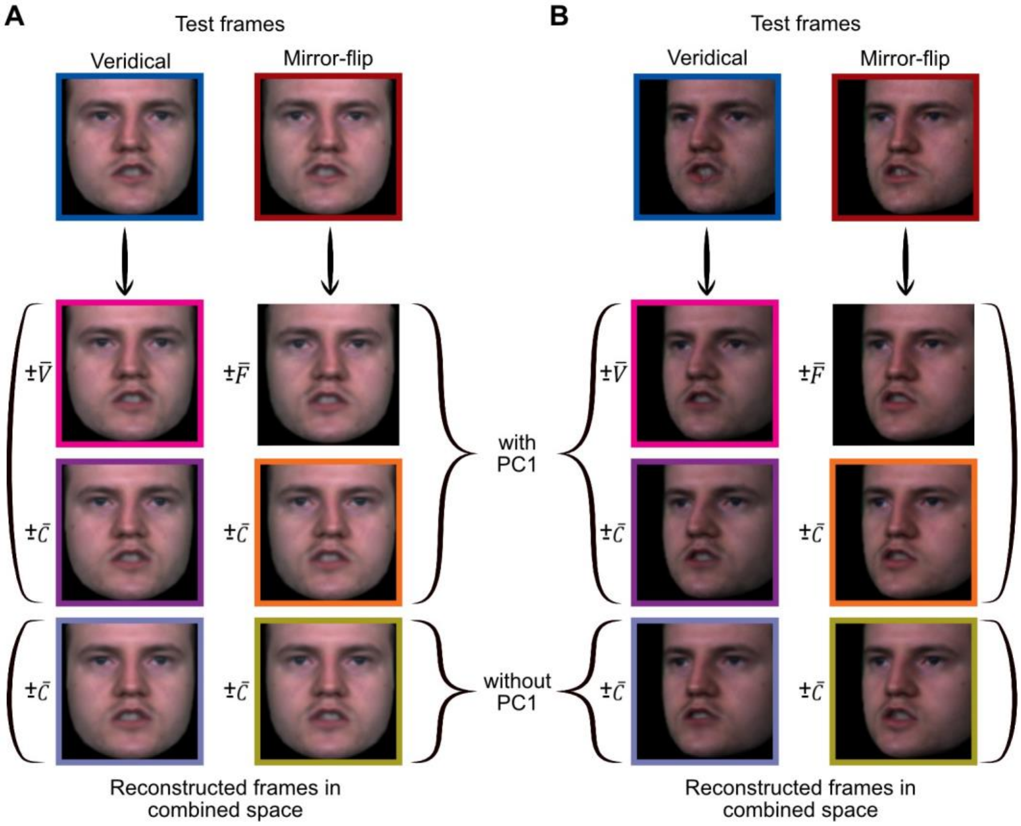


Figure 5.10. Reconstructed frame from the combined mirror space. Reconstructions of veridical and mirror-flipped frames in the combined space with and without PC1. The top row shows the veridical (blue) and mirror-flipped (red) frames for frontal (**A**) and left-facing (**B**) views. The row second from top shows the reconstructions of the veridical (pink, left) and mirror-flipped (no border, right) frames created by subtracting and re-adding $\bar{V}$ and $\bar{F}$ respectively. The bottom row and second from bottom show reconstructions of the veridical (purple + lavender) and mirror-flipped (orange + green) frames made by subtracting and re-adding $\bar{C}$ with (purple + orange) and without (lavender + green) PC1 included in the combined space. The coloured borders correspond to the colours used describe the methods in Figure 5.5 and Figure 5.12 and in the results in Figure 5.11 and Figure 5.13.

There was a main effect of reconstruction method across all views, features, and measures (all F(3,13347) > 4800, p < .001, η_p^2 > .52). Reconstructions were worse from the combined space than the veridical space (all p < .001) yet the differences were comparatively small (see the blue, pink, and purple bars in Figure 5.11). For instance, the difference in ED in image similarity for frontal frames between reconstructions in the veridical space and in the combined space using either method ($\pm\bar{V}$, $\pm\bar{C}$) was smaller (both δ = 55.31) than the reconstructions made from the flipped space (δ = 374.79). Reconstructions using the combined space were also significantly better than those using the flipped space, with lower EDs and higher Z-values (all p < .001). The methods used for the combined space ($\pm\bar{V}$, $\pm\bar{C}$) did not significantly differ (all p > .02) after Bonferroni-correction (α = .004).

We again assessed the loading similarity of the reconstructed deviation vectors relative to the veridical frames in the veridical space. $\bar{V}$ was subtracted from all reconstructions so that the deviation vectors would be relative to the origin of the veridical space. The loading similarity to the veridical projections was consistently better (lower ED and higher Z) for reconstructions made using the combined space than those made in the flipped space (all p < .001). In fact, no frames had a higher loading similarity when reconstructed in the flipped space compared to the combined space. There was no difference in the loading similarities between methods ($\pm\bar{V}$, $\pm\bar{C}$) for reconstructing in the combined space (all p > .042, α = .008).

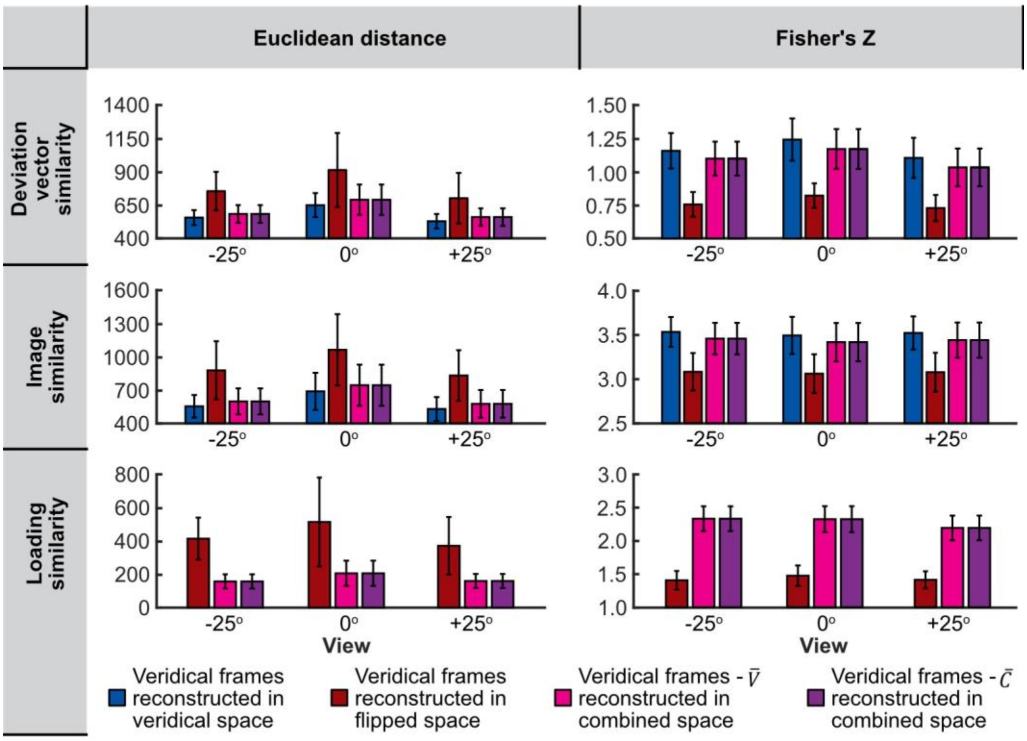


Figure 5.11. Reconstruction accuracy of veridical and mirror-flipped frames in the combined space.

Reconstruction accuracy (ED and Fisher's Z) for veridical frames reconstructed in the veridical (blue), the mirror-flipped (red) and the combined space after removing and re-adding $\bar{V}$ (pink) and $\bar{C}$ (purple). Deviation vector and frame similarity is measured relative to the veridical frames. Loading similarity is measured relative to the loadings of the veridical frames in the veridical space. The colour code corresponds to the methods in Figure 5.5 and the example reconstructions in Figure 5.10. Error bars show ±1SD.

As the first component seems to capture the structural asymmetries, we next assessed reconstructions made without PC1. Looking at the reconstructions (lavender + green in Figure 5.10) it is much more challenging to see the structural asymmetries. The nose is straight, and the mole appears faintly on both cheeks. Despite this, asymmetric motion was somewhat recoverable, suggesting asymmetries in the facial motion were captured by other components, indeed asymmetries are visible in other components particularly for the non-frontal views.

To test the importance of PC1 we performed analyses to assess 1) if reconstructions made without PC1 were significantly worse than those with PC1, and 2) whether the reconstructions of the veridical and flipped frames were more similar when PC1 was removed than when included. We assessed these effects using two analyses. The methods are shown in Figure 5.12. We projected the veridical and flipped frames into the combined space having first subtracted $\bar{C}$, and reconstructed the deviation vectors with and without PC1. To reconstruct the frames, we added $\bar{C}$ and de-warped the images. We compared the image and deviation vector similarity as well as the loading similarity in the veridical space using one-way within-subjects ANOVAs to compare the reconstruction methods. To test the loading similarity, $\bar{V}$ was again subtracted, relating the deviation vectors to the origin of the veridical space. In this first analysis, measures of reconstruction were all relative to the images, deviation vectors and loadings of the veridical frames.

In this analysis, there was a main effect of reconstruction method for all views, features and measures (all $F(3, 13347) > 13{,}600$, $p < .001$, $\eta_p^2 > .75$). Reconstructions of the veridical frames were consistently worse without, compared to with, PC1 (all $p < .001$), yet they were still significantly more similar to the veridical frames than reconstructions of the flipped frames (all $p < .001$). But, removing PC1 made the reconstructions of the flipped frames significantly more similar to the veridical frames (all p < .001), suggesting that the veridical and mirror-flipped reconstructions were significantly more similar to each other when PC1 was omitted.

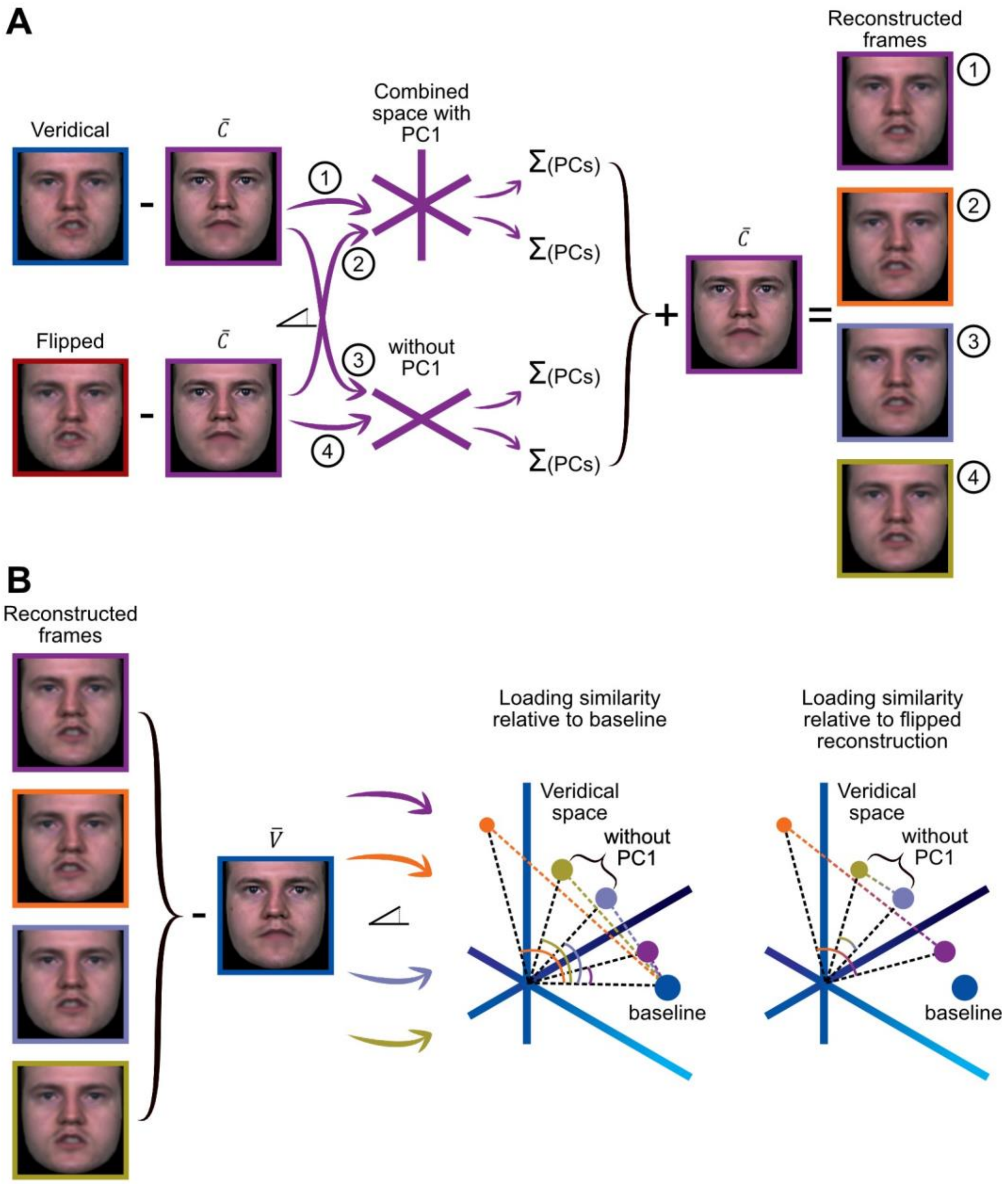


Figure caption on next page

Figure 5.12. Projection and reconstruction in the combined space with and without PC1.
(**A**) Veridical (1+3) and mirror-flipped (2+4) frames were projected into the combined space with (1+2) and without (3+4) PC1. The weighted components were then summed to reconstruct deviation vectors and $\bar{C}$ was added to reconstruct the frames. (**B**) Veridical and mirror-flipped reconstructions were projected into the veridical space and either compared to baseline (the loadings of the veridical frames) or to each other. Images show actual frames and reconstructions. Projections into the veridical space in (**B**) illustrate the methods and predicted results, not actual data. The colour code corresponds with the results in Figure 5.13 and reconstructions in Figure 5.10.

To test this further, a second analysis directly compared the image, deviation vector and loading similarities calculated between the veridical and mirror-flipped reconstructions, with and without PC1 (see Figure 5.12b). This is contrast to the previous comparisons in which similarity was calculated relative to the baseline (the actual veridical frames). This more direct comparison confirmed that the reconstructions of veridical and mirror-flipped frames from the combined space were more similar to each other when PC1 was removed (all $p < .001$).

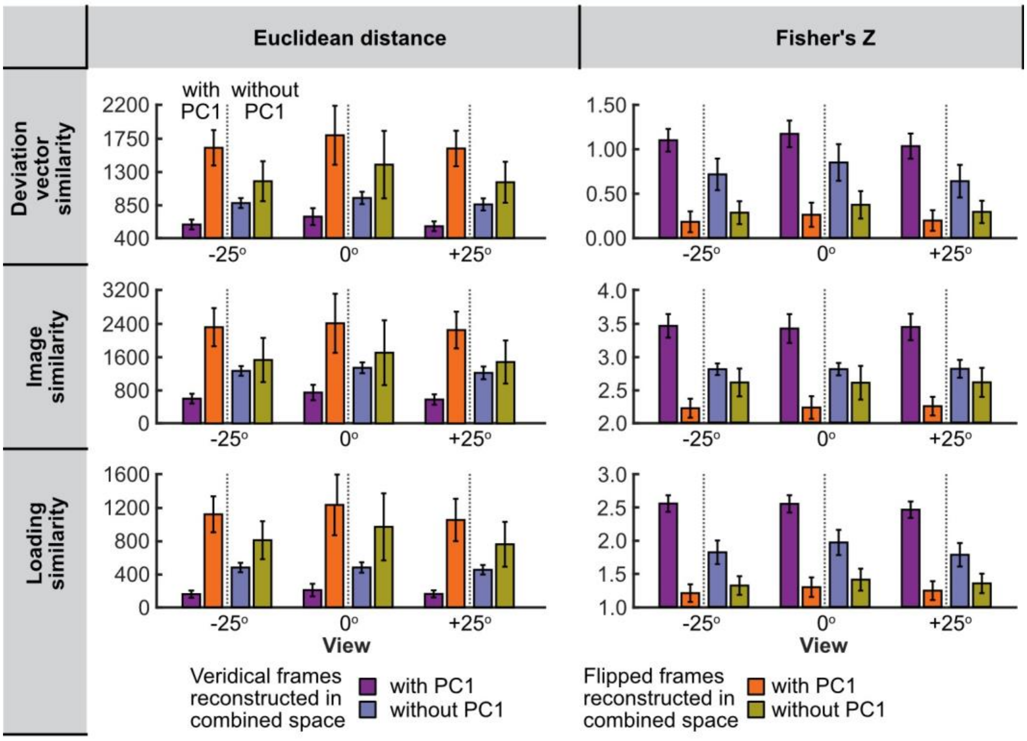


Figure 5.13. Reconstruction accuracy of veridical and mirror-flipped frames in the combined space with and without PC1. Reconstruction accuracy (ED and Fisher's Z) for veridical (purple, lavender) and flipped (orange, green) frames reconstructed in the combined space with (purple, orange) and without (lavender, green) PC1. Deviation vector and frame similarity was measured relative to the veridical frames. Loading similarity was measured relative to the loadings of the veridical frames in the veridical space. The colour code corresponds to the methods in Figure 5.12 and reconstructions in Figure 5.10. Error bars show ±1SD.

5.5 Discussion

The observed responses to mirror viewpoints within face-selective regions of cortex has led to suggestions that mirror views are a stepping stone to view-invariance (e.g., Meyers et al., 2015; Rogers & Andrews, 2022), yet the underlying mechanism is unclear. This chapter therefore explored a couple of ways that mirror views might be represented within a 2D face space representation and presented a method that can compress the representation of mirror views within our multi-view face space models. It discussed three theoretical models in the introduction and used two experimental models to test their plausibility.

The first theoretical model discussed contained view-specific spaces learned for each hemi-view. New information was projected into both the spaces corresponding to the input view and to the mirror viewpoint. This is conceptually similar to pooling across visual fields (e.g., see Corballis & Beale, 2020; Gross et al., 1977). The second related, albeit unlikely, model contained spaces which were only learned from one hemi-view, with the other hemi-view ignored. After learning, both hemi-views would be projected into this space for subsequent processing. Neither model collapsed across mirror views *per se*, but in both there was an assumption that information from one hemi-view should be retrievable from the other hemi-view and thus that facial motion needs to be sufficiently symmetrical.

In the first experimental model, reconstructions made by projecting frames into the mirror-opposite space were found to be significantly worse than those made from the hemi-view's own space. From visually comparing reconstructions and measures of reconstruction accuracy, we found that facial motion is asymmetric, consistent with prior results (Graves et al., 1982; Jordan & Thomas, 2007; Nicholls & Searle, 2006) and that it is challenging to reconstruct these asymmetries in the opposite view's space. This casts doubt on the efficacy on the first two models discussed, especially for the model which only learns from one hemi-view.

In the second experimental model we then assessed the plausibility of encoding a collapsed version of both hemi-views, where one view is mirror-flipped, and both views are encoded together in 2D space. We found that we could encode and reconstruct motion well from both hemi-views, and that reconstructions were better from this combined space than from the mirror-flipped space.

The first component reflected major structural asymmetries, with subsequent components being symmetrical in structure. The subsequent components coded both symmetric and asymmetric dynamic information. The coding of symmetric motion in the combined space may reduce redundancies over encoding the same symmetrical actions in both hemi-views separately. Despite coding two hemi-views instead of one within the same number of components, the combined space did a sufficient job of capturing both symmetrical and asymmetrical information.

While the first component for this actor captures structural asymmetries well, this may not always be the case. For other actors with more symmetrical faces, or for videos captured under more variable conditions we may find different results. For example, the first component may instead reflect a change in lighting (e.g., see Burton et al., 2016) with subsequent components coding structural asymmetries.

For the actor shown, the lip separation is greater on the actor's left side than right side. This is at odds with evidence that, *on average*, lip separation and movement is greater on the right side of the face, particularly for a right-handed male (Graves et al., 1982; Jordan & Thomas, 2007; Nicholls & Searle, 2006), but this is not unexpected given individual variation and bears no impact of the efficacy of the model.

In relation to the multi-view face space model in the previous chapter, our results support the possibility that mirror viewpoints can be collapsed prior to creating the final multi-view representation. Rather than having, say, 9 separate spaces or slots covering -90^{o} to +90^{o} (or 10 if also mirror-flipping frontal), one could simply have 5 spaces from 0^{o} to 90^{o}. Note there is ambiguity over the frontal face. Flipping 0^{o} seems somewhat illogical, but if you mirror-flip the other views then where is the boundary? Also, in the current digital age, frontal faces are often mirror-reflected by the front-facing cameras on phones and laptops. Nevertheless, the results here suggest that it would be plausible for both hemi-views to be represented in one space, providing a route to forming a more compact multi-view face space.

This work has shown how mirror views might be collapsed within a 2D, multi-view face representation, yet there is still much research to be done. For instance, behavioural experiments would be needed to establish if the brain does collapse across mirror views. Evidence for this could come from ensemble coding. Participants incorrectly identify the average of four seen images as a previously seen image itself (Davis et al., 2021) showing evidence that the brain automatically extracts averages, even when stimuli are presented in an atypical manner, such as when inverted (Davis et al., 2021). If the brain collapses across hemi-views, then the average across the hemi-views should also be mistakenly identified as a previously seen image, although perhaps only under challenging task conditions.

Furthermore, it would need to be determined if collapsing across mirror views like this helps in achieving view-invariance, or if there is another step or process necessary. It could be the case that collapsing across mirror views populates the view-specific spaces faster than representing each hemi-view separately, thus making it quicker to formulate the view-invariant, multi-view representation. It may also be easier to construct this multi-view representation with half the number of view-specific slots.

In summary, this chapter addressed a couple of ways in which hemi-views of faces in motion might be collated and represented within a 2D face space representation and the multi-view face space model presented in the previous chapter. From our results it seems that both hemi-views are necessary to suitably encode asymmetric facial motion, but that it is possible to collapse both hemi-views into a single face space representation. These methods potentially provide a route to simplifying the transition from view-dependence to view-invariance.

5.6 Supplementary materials

5.6.1 Supplementary videos

Supplementary videos can be found using the link below:

https://www.dropbox.com/scl/fo/kaqognq18e0a5vira9syo/h?rlkey=m0k5mrhvglk69rpjdy92lk5ha&dl=0

Within the Supplementary Materials folder, navigate to 'Supplementary_Videos/Chapter_5_mirrored_motion'. This folder contains videos from both models. Within each model there are folders containing 'reconstructions' and 'pc_videos' (videos of the principal components). There are Word documents ('info.docx') provided throughout to provide information about the videos.

Chapter 6 Investigating the presence of prototypical views

6.1 Preface

One of the key assumptions in the multi-view PCA work outlined in Chapter 4 is that a few prototypical views are a represented at the view-dependent stage rather than every possible viewpoint. In the model these were assumed to be at 0^{o}, 22.5^{o}, 45^{o}, 67.5^{o} and 90^{o}. In the current chapter we wanted to investigate which views the brain might use to form these representations, both to further understand the face processing system and to guide future model development.

6.2 Introduction

When talking to a friend, loved one, or even a stranger, we can often process emotional expression, health, age, identity, gender and more regardless of viewpoint. While the process for achieving view-invariance is not fully understood, it is widely recognised that the hierarchy to invariance starts with a level of view-dependence (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Meyers et al., 2015; Perrett et al., 1991; G. Wang et al., 1996, 1998). Unlike in macaques, little research has been conducted in humans to assess which views might be involved in this view-dependent stage. Understanding which views might be preferentially represented has many important applications, from improving computational models such as in Chapter 4 to guiding which viewpoints should be used in police lineups.

Early research in macaques (Perrett et al., 1988, 1991; G. Wang et al., 1996, 1998) showed evidence of neurons preferentially tuned to different prototypical views, including frontal (0°), profile (90°/270°), back, tilted upwards and tilted downwards views, with fewer specifically tuned to 45° and even fewer to intermediate views (22.5°, Perrett et al., 1991). View-selective neurons are seen in partially overlapping clusters (Perrett et al., 1988) with neighbouring views occupying neighbouring regions of cortex (G. Wang et al., 1996, 1998). Even view-selective neurons however show large tuning widths, responding to a broad range of views (Perrett et al., 1991). It is likely that the combination of a few prototypical views and broad tuning curves is computationally optimal for achieving view-invariance compared to representing every possible view.

Following learning of novel objects, Logothetis and Pauls (1995) found that monkeys could often interpolate between learned views up to 120° apart, and found that that different, view-selective neurons tended to be preferentially tuned to views ~40-50° apart. Although this is in the context of object processing, it is consistent with the evidence of neurons preferentially tuned to frontal (0°), profile (90°) and 45° views of faces (Perrett et al., 1991).

While single unit recordings provide invaluable insights into the activity of single neurons, they are limited in by how many neurons they can record from; other viewpoints may be represented, but neurons representing those views are not sampled. Perrett and colleagues

(1991) for instance only recorded from 120 neurons in the polysensory area of the macaque STS. Logothetis and Paul (1995) recorded from over 700 neurons but were testing responses to objects. The number of views necessary to encode dynamic faces may be more than for coding objects. Moreover, view-invariance may be achieved differently in humans and macaques as their face space representations have been suggested to differ (Parr et al., 2012), so it is important to investigate which views are preferentially represented in humans.

While some research supports a 2D interpolation account (H. H. Bülthoff & Edelman, 1992) of face processing (W. Chen & Liu, 2009; Y. Lee et al., 2006; Liu et al., 2009; Schwaninger et al., 2007; Wallraven et al., 2002), it is not yet clear which views contribute to this interpolation system. Lee and colleagues (2006) found evidence suggesting a categorical representations of view, however they only tested a limited range of viewpoints (0°, 6.7°, 13.3°, 20°). They did not test viewpoints outside of this range to establish how many categories of view there are or whether the group size changes with distance from 0°, nor did they determine where between 6.7° and 13.3° the boundary occurs. This grouping has also not been seen in other research with slightly different viewpoints (Swystun & Logan, 2019), although see a critique of this paper below.

Behavioural evidence in humans for preferential views is mixed. Some studies show a ¾ view (~45°) advantage (Marotta et al., 2002; O'Toole et al., 1998; Troje & Bülthoff, 1996; Van der Linde & Watson,

2010) over both frontal and profile, suggesting that the ¾ view is preferentially represented. In contrast, others show a frontal view advantage (Favelle et al., 2017; H. Hill et al., 1997), and some show no difference between these views (Carbon & Leder, 2006; Favelle & Palmisano, 2018). Sama and colleagues (2019) found a frontal advantage for ensemble coding of viewpoint, but no effect of view on the ensemble coding of identity.

Swystun and Logan (2019) also found a frontal advantage over more intermediate views (5°, 10° and 20°) which is compatible with neural preferences for frontal views over intermediate views (Perrett et al., 1991). However, the facial features always appeared to be of a front-facing image, placed into their respective positions on a rotated face, leading to an uneasy percept. More work is needed to verify their findings.

The apparent preference for 45° may stem from having a neural population preferential to this view, from the overlap between broad tuning curves for frontal and profile views, or most likely, a combination of both. Interpolation alone between neurons tuned to frontal and profile is likely insufficient to support a behavioural advantage. As view-selective neurons on average have a full width at half-maximum (FWHM) of 60° (Perrett et al., 1991), their firing rate will have decreased by ~80% by 45°. Even when summed this is only 40% compared to frontal or profile, although, some neurons are more broadly tuned. In macaques, Perrett and colleagues (1991) also found far fewer neurons tuned to 45^o than frontal or profile, suggesting that the behavioural advantage for 45^o cannot solely be explained by neurons preferentially selective to this view. However, in more anterior and dorsal portions of the STS, De Souza and colleagues (2005) found more neurons sensitive to oblique views than frontal or profile, potentially explaining the behavioural results.

While many studies have looked at behavioural preferences to different viewpoints, this has either been on quite a coarse scale, for example ≥15^o between viewpoint (Favelle & Palmisano, 2018; Marotta et al., 2002; O'Toole et al., 1998; Troje & Bülthoff, 1996; Van der Linde & Watson, 2010), or on a finer scale but over a smaller range (Y. Lee et al., 2006). This is likely because of the challenges of investigating a finer scale over a larger range, however, to build a comprehensive picture of whether certain viewpoints are better represented than others, it is a necessary task. For context, 19 conditions are required to transition from frontal (0^o) to profile (90^o) in 5^o increments in only one direction. Kietzman and colleagues (2017) have a dataset of EEG responses to stimuli at these 19 viewpoints, but they did not report any results regarding, for example, response amplitudes to the different views. Such an analysis would help provide evidence for prototypical views.

Problematically, the task may confound this research. For example, Van der Linde and Watson (2010) showed a frontal view advantage when participants did not need to match identity across views, but a ¾ advantage when they did. This behavioural change may indicate a change in the neurons used for the task, with same-view matching recruiting more posterior neurons (Perrett et al., 1991) and cross-view matching more anterior neurons (De Souza et al., 2005). This may indicate that the ¾ view might be especially informative for achieving view-invariance. Three-quarter views may provide the best visibility of transformation points, provide the most information about 3D shape or contain the most diagnostic information about identity (H. Hill et al., 1997; Marotta et al., 2002; Troje & Bülthoff, 1996). Regardless of the exact reason for the ¾ preference for cross-view matching, it is clear that the task might confound research into which viewpoints are best represented, especially if trying to understand the first, view-dependent stage of the hierarchy.

One way to avoid this confound is to make the task as perceptual as possible, rather than cognitive. Like many functions, face processing involves a hierarchy of processes, starting with perceptual processing of low-level properties of the face, such as certain colours and shapes, and eventually ending with higher-level properties such as gender, identity and expression recognition, and impression formation. Recently, Diane Beck and colleagues (Caddigan et al., 2017; Center et al., 2022; P-L. Yang & Beck, 2021, 2022, 2023) have provided some excellent insight into higher-level distinctions for faces and objects using lower-level perceptual processes. Rather than asking participants to perform identity judgements for example, requiring a more cognitive judgement, they were asked to discriminate the intact image from either a phase scrambled (Caddigan et al., 2017) or morphed version of that image (Center et al., 2022), requiring a more perceptual judgement. Morphing using Diffeomorph (Stojanoski & Cusack, 2014) retains most of the low-level properties whilst highly distorting the stimuli.

In scene and object processing, images rated as highly representative of a particular category are better discriminated from scrambled/warped images than those not rated as highly representative (Caddigan et al., 2017; P-L. Yang & Beck, 2022). These stimuli were also more easily discerned by support vector machines, and lead to better learning in those models (P-L. Yang & Beck, 2022). Participants were also faster to discriminate objects from scrambled versions in typical orientations than atypical orientations when rotated in depth, such as viewing a lamp from the side rather than underneath (Center et al., 2022). The results further suggested that particularly informative viewpoints lead to faster discrimination, not just the amount of experience with the viewpoint. This again argues for a 2D interpolation account based on prototypical views.

In the context of faces, Yang and Beck (2021, 2023) found that upright and famous faces are more easily discriminated from scrambled/warped images than inverted and unfamiliar faces. Furthermore, the advantage for upright faces was stronger for famous than unfamiliar faces. This interaction was not seen for brand logos, suggesting the benefit of famous upright faces reflects a face-specific process (P-L. Yang & Beck, 2021, 2023). These results suggest that at the purely perceptual level, there is a more efficient representation for familiar, famous faces than unfamiliar faces. To reiterate, participants did not have to recognise the faces, they simply had to detect the veridical images.

Here, we used similar methods to assess for prototypical horizontal viewpoints. We hypothesised based on the recordings in macaques (De Souza et al., 2005; Perrett et al., 1991) and behavioural results (Marotta et al., 2002; O'Toole et al., 1998; Troje & Bülthoff, 1996; Van der Linde & Watson, 2010) that thresholds for distinguishing veridical and morphed stimuli would be smallest for 0^{o}, 45^{o} and 90^{o} viewpoints and largest for intermediate viewpoints.

To minimise the influence of lower-level confounds, we also included blocks of picture plane inverted stimuli. We firstly hypothesised that thresholds would be larger for inverted stimuli. Secondly, we hypothesised that any increases in the thresholds for certain views would be larger for inverted stimuli than upright stimuli. We assumed that frontal, ¾ and profile views would be well represented and could be processed efficiently when upright and reasonably efficiently when inverted. If intermediate views are presented, we reasoned that interpolation would be more efficient for upright faces than inverted faces, giving rise to more prominent inversion effects for these intermediate views.

6.3 Materials and methods

6.3.1 Participants

Ten healthy volunteers with normal or corrected-to-normal vision were recruited for this study. Participants were aged between 18 and 33 years old (mean = 26 years, 6 months, SD = 5 years, 6 months). Seven were female, 3 male.

From self-reports, 7 participants were very familiar with the researcher prior to the study, 2 were completely unfamiliar, and 1 was only slightly familiar. Familiar participants reported that they were familiar with and could recognise the researcher from frontal, ¾ and profile views.

No other demographic details were collected. The sample was a mix of postgraduate research students and staff from the School of Psychology at the University of Nottingham. Fellow PhD students who were familiar with the researcher were recruited through a combination of convenience and snowball sampling. Undergraduates with no familiarity with the researcher were recruited through the School's Research Participation Scheme. The study was approved by the School's ethics committee.

6.3.2 Apparatus

The experiment was built in MATLAB version 9.5 (R2018b) using the Psychophysics Toolbox extensions (Psychtoolbox-3 version 3.0.17, Brainard, 1997; Kleiner, 2007; Pelli, 1997).

Due to restrictions in laboratory availability, the experiment was conducted across two laboratories. The same model of CRT monitor was used in both (Mitsubishi DPlus 230SB, 1024 x 768 pixels), one running Windows 7, with a screen refresh rate of 85Hz, the other, Ubuntu 18.04 at 120Hz. Viewing distance was 100cm in both instances. Head position was maintained using a head and chin rest. No obvious differences in the data were observed between the two laboratories.

6.3.3 Stimuli

Due to the length of the experiment and the number of potentially confounding factors, we focused on controlling effects of identity by using one actor. Due to being in the height of the Covid-19 pandemic, stimuli were captured of the researcher (RE). The actor maintained a neutral expression throughout.

Photographs were taken from 0^{o} (frontal) to 90^{o} (left profile) in 5^{o} increments (19 views total) on a Huawei P20 Pro. Photographs were taken 110cm away from the actor, with the camera positioned using a semi-circular camera rig covering a range of 180^{o} in total (see Figure 6.1). Illumination was provided through the ceiling lights and two diffuse lamps (InterFit F5, 50 x 69 cm), positioned at 20^{o} and 70^{o} yaw from frontal, with the centre of the lamps 20^{o} above the camera position, at approximately 128cm from the actor.

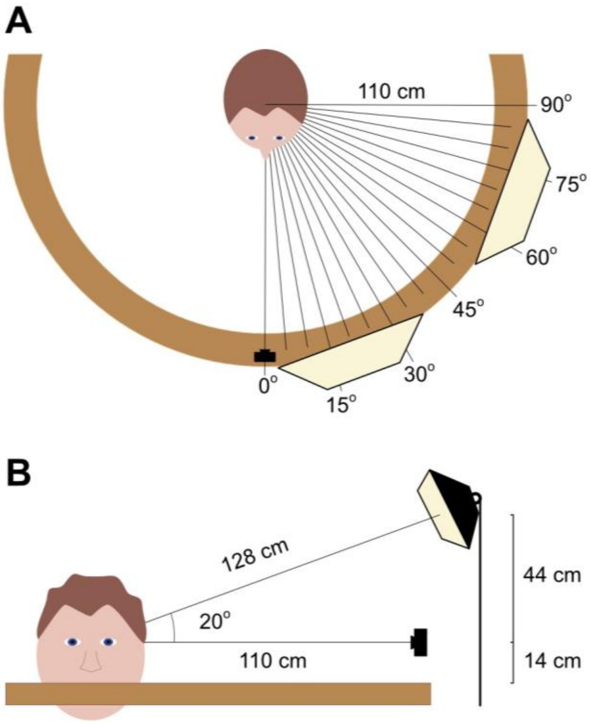


Figure 6.1. Stimulus capture

(**A**) The semi-circular camera rig, with lines reflecting the angles photographs were taken from. Pictures were taken 110cm from the actor. The trapeziums show the position of the diffuse lights in azimuth. (**B**) Distance and angle of the camera and diffuse lights relative to the actor in elevation. Vertical head position is approximate.

The backgrounds were removed from the stimuli by hand to keep only the face, hair and neck. Stimuli were then morphed using Diffeomorph (Stojanoski & Cusack, 2014). Each of the 19 stimuli were warped to generate 20 unique sets of 80 warp levels. Twenty unique sets were generated to reduce confounding effects of specific warping to each viewpoint (see Figure 6.2a). Each set was used twice per viewpoint in each block, in a random order. On average, stimuli (not including the backgrounds) occupied ~8° of visual angle.

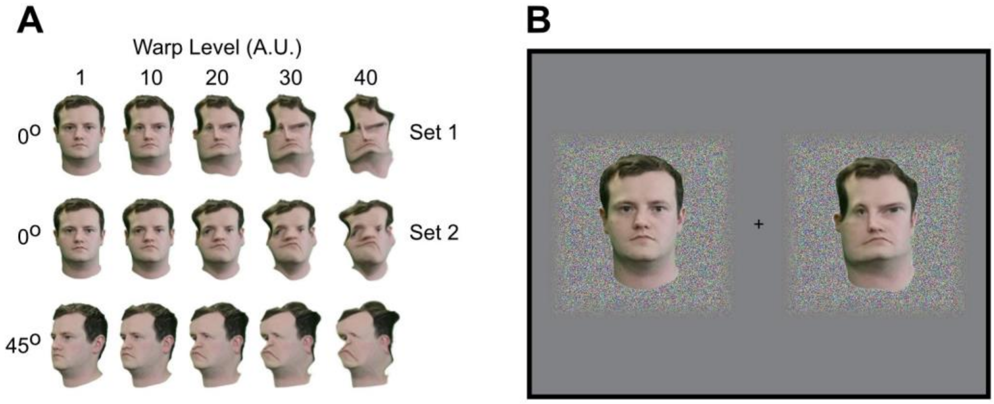


Figure 6.2. Example stimuli and trial in the prototypical views experiment.
(**A**) Example stimuli. The top two rows show example stimuli made by warping the same image twice, creating two different sets of warps. The bottom row shows some examples of warping the 45° stimulus. A warp level of 1 is the intact, unedited image. (**B**) An example trial. Participants fixate on the central fixation cross, a stimulus appears on either side for 300ms, and participants have to determine which image (left or right) was not warped.

6.3.4 Task

The task consisted of a psychophysics experiment using a staircase method to test whether there is a difference in thresholds for correctly identifying the non-morphed image across different viewpoints. On each trial, participants were presented with 2 stimuli simultaneously, one either side of a central fixation cross. One stimulus was an unedited image (veridical), the other a morph of that image. Participants had to identify the veridical stimulus using the left and right arrow keys. As the trials progressed, the warp level (arbitrary unit) decreased using a Quest staircase procedure (Watson & Pelli, 1983), starting at a warp level of 46. This differed from the methods by Yang and Beck (2021, 2023) who instead used extremely morphed stimuli and modulated the presentation duration. The reason for not modulating the duration here was because a suitable starting duration for upright stimuli was far too short for inverted stimuli, but a longer initial duration decreased the chance of reaching threshold for upright stimuli.

Three different viewpoints were presented within each block to reduce the reliance on certain external contours and internal configurations. Each viewpoint was presented in 3 different blocks (within a given direction x inversion condition), paired with different viewpoints in each. As 3 of the 19 viewpoints were presented in each block, and each viewpoint was presented 3 times, this gave 19 blocks. All blocks were presented four times: upright and inverted x original (rotating from frontal to left profile) and mirror-flipped (frontal to ‘right’ profile) directions. This gave a total of 76 blocks (19 view sets x 2 orientations x 2 directions). Within each block, 40 trials were presented per viewpoint, giving 120 trials per block presented in a random order, and 9120 across the whole experiment. Each block took on average 3 minutes 40 seconds (SD = 37 seconds). The experiment took approximately 8 hours in total, split into several sessions. The division of blocks was up to participants, but they were encouraged to complete 8 sessions of ~10 blocks.

Participants maintained fixation on the black central fixation cross throughout. Stimuli were presented either side of this cross for 300ms. Responses could only be made after the stimulus was removed, indicated by the fixation cross turning white until response.

Response time was unlimited, and the next trial began 1s after response. Reaction times were recorded but not analysed as participants were instructed to focus on accuracy.

Stimulus size and position jittered slightly across trials to reduce retinotopic adaptation. Horizontal and vertical position was randomly jittered by 0, 5, 10 or 15 pixels in each direction. Size was modulated by expanding or contracting the stimuli by 0, 10, 20 or 30 pixels in each direction. This was applied to the whole square image, so aspect ratios were unaffected. Stimuli were presented on coloured backgrounds made using random noise in each of the RGB channels, which changed at the onset of each new stimulus. Masks were feathered into the background and their size and position remained constant.

6.3.5 Inter-feature distance

It is possible that the results may be predicted based on distances between features. To measure the inter-feature distances, up to 82 landmarks were positioned on each of the 19 veridical stimuli using InterFace (Kramer et al., 2017), depending on landmark visibility. In some cases, the landmark itself was not visible but the marker was still used. For instance, the landmark for the actor's right tragus was used to mark the edge of the cheek at the same height when the tragus was occluded. While the distances may generally be proportional to the cosine of the angle from frontal, they may also deviate from this. For instance, a measure of total visible width will likely peak at the view where the nose begins to extend the external contour past the cheek.

The features were initially collapsed into 8 horizontal distances. (1) The whole visible range of the face (excluding the ears, hair and back of the head). (2) The whole visible size of the mouth and (3) the distance from the centre of the top lip to the actor's left corner of the mouth. (4) The whole visible size of the nose. The distance between the tip of the nose and (5) the actor's left flange, (6) the left pupil and (7) the left tragus. (8) The inter-pupil distance, although note that the position of the actor's right pupil was estimated once occluded.

6.4 Results

To summarise the primary results, we found that upright faces were processed more efficiently than inverted faces, and that in general processing efficiency reduced as the distance from frontal increased. The inversion effect linearly increased with increasing angular distance from frontal but was proportional to the threshold for upright stimuli; the threshold for inverted stimuli was on average 1.48x the threshold for upright stimuli, irrespective of viewpoint.

To investigate the effect of viewpoint, we collapsed across direction (left/right) giving 6 runs in total per condition. Trials from each of these 6 runs were collated and used to fit one Weibell function per condition (240 trials). Trials where participants were assumed to have made a response error by responding incorrectly to a high warp level were removed. These response errors were determined by listing the warp levels for the incorrect trials for each condition and removing any trials where the warp level was >3SD away from the mean. The threshold at 75% accuracy was then calculated (see Figure 6.3). This was performed at subject-level.

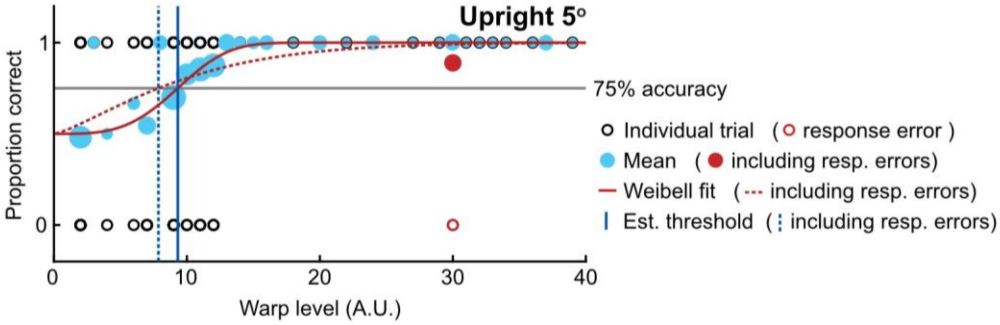


Figure 6.3. Threshold calculation.

Demonstration of threshold calculation for upright stimuli at 5° for one participant. All trials for a condition were collated together, coded as either correct (1) or incorrect (0). Individual trials are shown by empty circles (red = response error). The proportion correct for each warp level is shown by filled circles, with the area being proportional to the number of trials for that warp level. The dark red and light blue circles show the proportion correct including and excluding response errors respectively. The horizontal grey line shows the 75% accuracy threshold. The dotted red curve and blue vertical show the Weibell fit and estimated threshold respectively including response errors. The solid red curve and blue vertical show the Weibell fit and estimated threshold respectively after excluding response errors.

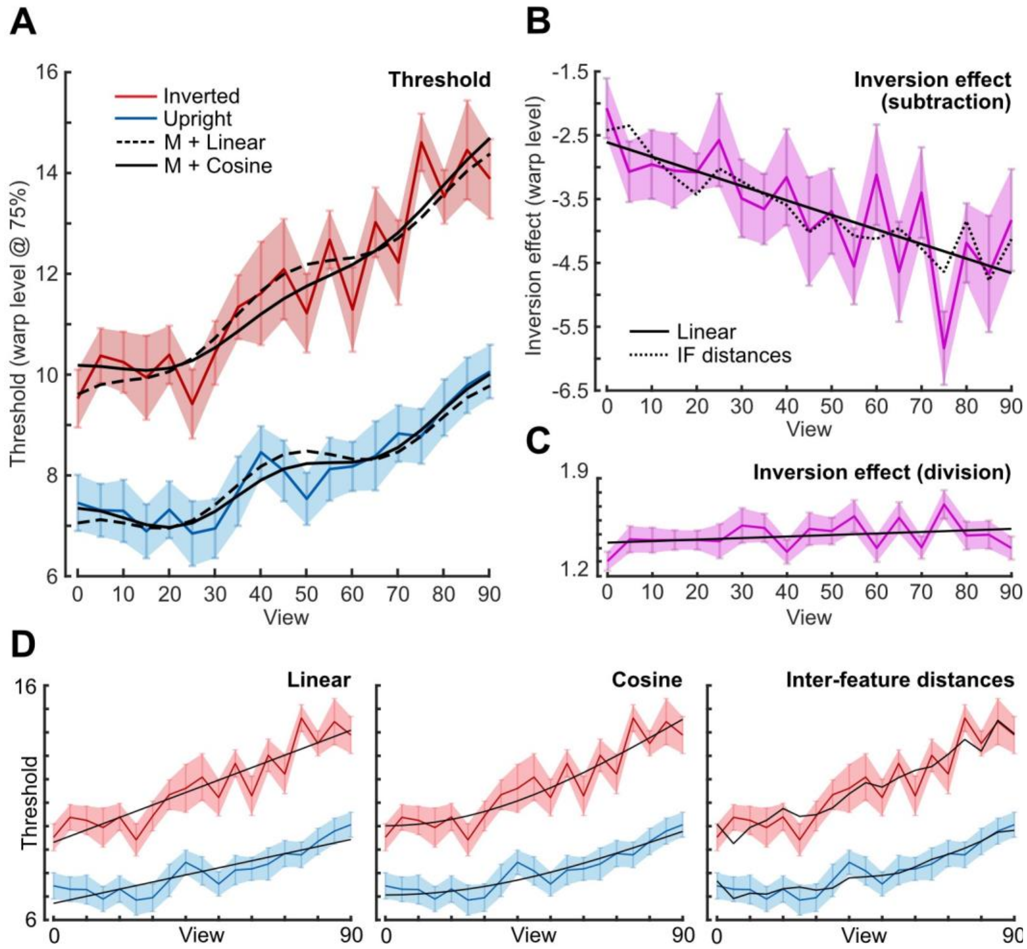


Figure 6.4. Average thresholds for upright and inverted stimuli. (**A**) The average warp level required for participants to distinguish a warped from a veridical image at 75% accuracy for upright (blue) and inverted (red) stimuli as a function of viewpoint. The solid black lines show the fit of the '*M + cosine*' models, the dashed black lines show the fit of the '*M + linear*' models. (**B&C**) The average inversion effect for each viewpoint, calculated by (**B**) subtracting the thresholds for the inverted stimuli from the thresholds for upright stimuli and (**C**) dividing the threshold for inverted by that of upright. The solid black lines show the fit of the linear model. In **B** the dotted black line shows the fit of the inter-feature distance model. (**D**) The model fits for the data in **A**, with the linear (left), cosine (middle) and inter-feature distance (right) models. In all plots, the error bars/area show ±1 SEM of within-subjects variance.

Figure 6.4 shows the 75% accuracy thresholds for correctly identifying the intact upright and inverted stimuli averaged across participants. Other than an apparent inversion effect in most participants, and a general trend for an increase in threshold with angular distance from frontal, there were no other pronounced effects from looking at individual data separately. Group-level analyses were therefore performed.

As the thresholds for upright and inverted stimuli were approximately linear, a 2 x 19 ANOVA, with orientation (upright and inverted) and viewpoint (0^o, 5^o, … 90^o) was first performed to assess the main effect of inversion and to establish if there are linear trends in the threshold as a function of angular distance from frontal. The dependent variable was the warp level at the 75% threshold.

There was a significant inversion effect. Thresholds were significantly lower for upright stimuli than inverted stimuli ($F(1,9) = 56.60$, $p < .001$, $\eta_p^2 = 0.86$). There was also a significant effect of viewpoint ($F(18,162) = 25.61$, $p < .001$, $\eta_p^2 = 0.74$) and a significant interaction between orientation and viewpoint ($F(18,162) = 3.49$, $p < .001$, $\eta_p^2 = 0.28$).

As there was a significant interaction between viewpoint and inversion, the effect of viewpoint for upright and inverted faces was further analysed separately. To analyse the effect of viewpoint (0^o, 5^o, … 90^o), five pairs (upright / inverted) of models were considered: two that treated viewpoint linearly; two that treated viewpoint non-linearly based on the cosine of the angle from frontal; and one that used inter-feature distances. For each, a linear mixed model was performed with a fixed slope and random intercepts. The dependent variable was the estimated threshold.

For the two linear and cosine models, two options were considered. The first simply tried to predict the threshold based on the distance from frontal. The second options added another regressor based on the hypothesised predictions. We initially predicted that the thresholds would be lower for frontal faces, 45° and profile than intermediate views. From the plot it almost appears that the opposite trend was present, superimposed on top of a linear or cosine function, with an increase in threshold at 0°, 45° and 90°. The second regressor was therefore constructed based on a sinusoidal curve with the lowest values at 0°, 45° and 90°, and the highest values at 22.5° and 67.5°. This was constructed using the following prompt in MATLAB: "*m = 1 – cosd(8v)*" where *v* is the vector of views (0:5:90). This regressor formed an 'M' shape so for simplicity is referred to as 'M'.

For the models using the inter-feature distances, multi-collinearity was first dealt with by removing variables such that no correlation was greater than 0.8. Four distances remained: the total width visible; the nose tip to actor's left tragus; the nose tip to left pupil; and the total mouth visible. In the upright model, all 4 distances were significant when entered in separate models as sole regressors (all

$p < .006$). In the inverted model all were significant save for the distance between the nose tip and actor's left pupil ($p = .740$). The overall fit of the model was not significantly reduced by removing this distance ($\chi^2(1) = 3.39$, $p = .066$) so for consistency with the upright model, this distance was retained.

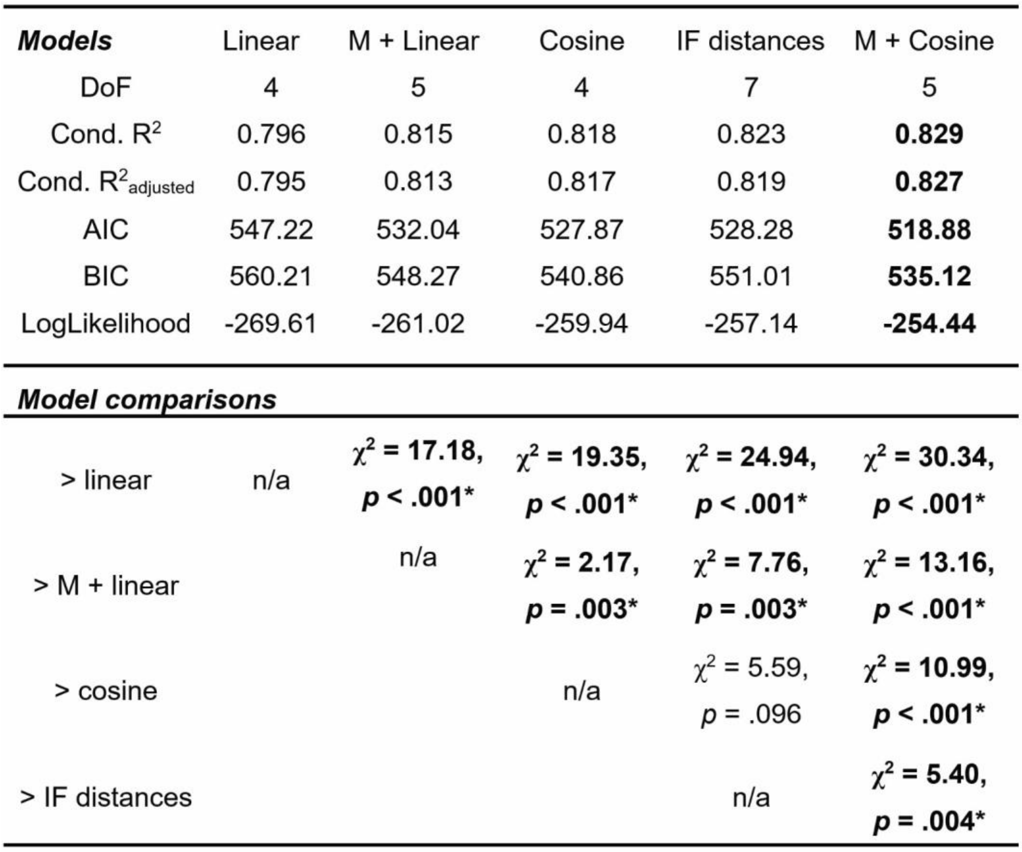
6.4.1 Upright stimuli

Firstly, the general effect of viewpoint was evaluated using the linear fit. The linear model showed that as the distance from frontal increased, the threshold increased (slope: $t(188) = 12.71$, unstandardized $\beta = 0.15$, $p < .001$, 95% CI = [0.13, 0.17]).

The model fits were then compared to see if a linear or non-linear function could best explain the data. Model comparisons were performed by comparing the explained variance, the Akaike and Bayes Information Criteria (AIC and BIC) and through Simulation Likelihood Ratio Tests (SLRTs) with 1000 simulations. Model comparisons are presented in Table 6.1 and the average fit for each model in Figure 6.4a.

The best model was the *M + cosine* model. It explained the most variance, had the lowest AIC and BIC and the highest log likelihood. It was also significantly better than all other models in the SLRTs. Both the cosine and M regressors were significant predictors (cosine: $t(187) = 14.31$, unstandardised $\beta = 2.67$, $p < .001$, 95% CI = [2.30, 3.03], M: $t(187) = -3.37$, unstandardised $\beta = -0.28$, $p < .001$, 95% CI =

[-0.44, -0.13]). As expected from viewing the plots, the sign of the coefficient for the M regressor was flipped compared to what was initially hypothesised.

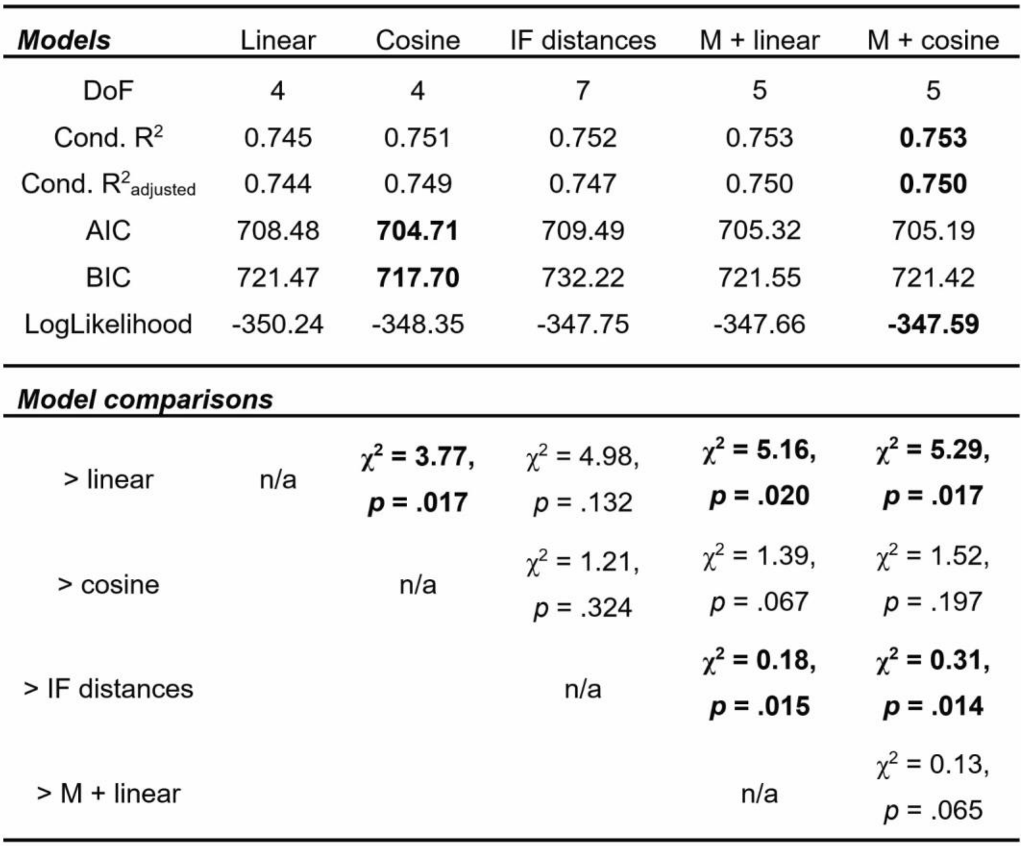


Models	Linear	M + Linear	Cosine	IF distances	M + Cosine
DoF	4	5	4	7	5
Cond. R^2	0.796	0.815	0.818	0.823	**0.829**
Cond. $R^2_{adjusted}$	0.795	0.813	0.817	0.819	**0.827**
AIC	547.22	532.04	527.87	528.28	**518.88**
BIC	560.21	548.27	540.86	551.01	**535.12**
LogLikelihood	-269.61	-261.02	-259.94	-257.14	**-254.44**
Model comparisons					
> linear	n/a	**χ^2 = 17.18,** ***p* < .001***	**χ^2 = 19.35,** ***p* < .001***	**χ^2 = 24.94,** ***p* < .001***	**χ^2 = 30.34,** ***p* < .001***
> M + linear		n/a	**χ^2 = 2.17,** ***p* = .003***	**χ^2 = 7.76,** ***p* = .003***	**χ^2 = 13.16,** ***p* < .001***
> cosine			n/a	χ^2 = 5.59, *p* = .096	**χ^2 = 10.99,** ***p* < .001***
> IF distances				n/a	**χ^2 = 5.40,** ***p* = .004***

Table 6.1. Model comparisons for upright stimuli.

Model accuracy and comparisons for the upright thresholds. In the top section, bold highlights the model with the best value. For AIC and BIC this is the lowest value, for log likelihood it is the highest. The bottom section shows model comparisons from Simulated Likelihood Ratio Tests. Bold = significant at *α* = .05, * = significant once corrected for 10 comparisons (*α* = .005). AIC = Akaike's Information Criteria, BIC = Bayes Information Criteria. Models were ordered based on the log likelihood.

While the model explained the data well, there is scope for improvement; the central peak in the threshold occurred slightly earlier than the model predicted at 40° rather than 45°, and had a sharper fall either side. Further work would need to replicate this finding before any fine tuning of the model would be valid.



Models	Linear	Cosine	IF distances	M + linear	M + cosine
DoF	4	4	7	5	5
Cond. R^2	0.745	0.751	0.752	0.753	**0.753**
Cond. $R^2_{adjusted}$	0.744	0.749	0.747	0.750	**0.750**
AIC	708.48	**704.71**	709.49	705.32	705.19
BIC	721.47	**717.70**	732.22	721.55	721.42
LogLikelihood	-350.24	-348.35	-347.75	-347.66	**-347.59**
Model comparisons					
> linear	n/a	**χ^2 = 3.77, *p* = .017**	χ^2 = 4.98, *p* = .132	**χ^2 = 5.16, *p* = .020**	**χ^2 = 5.29, *p* = .017**
> cosine		n/a	χ^2 = 1.21, *p* = .324	χ^2 = 1.39, *p* = .067	χ^2 = 1.52, *p* = .197
> IF distances			n/a	**χ^2 = 0.18, *p* = .015**	**χ^2 = 0.31, *p* = .014**
> M + linear				n/a	χ^2 = 0.13, *p* = .065

Table 6.2. Model comparisons for inverted stimuli.

Model accuracy and model comparisons for the inverted thresholds. In the top section, bold highlights the model with the best value. For AIC and BIC this is the lowest value, for log likelihood it is the highest. The bottom section shows model comparisons from Simulated Likelihood Ratio Tests. Bold indicates significance at *α* = .05. No comparisons survived Bonferroni correction (*α* = .005). AIC = Akaike's Information Criteria, BIC = Bayes Information Criteria.

6.4.2 Inverted stimuli

In general, the models for inverted stimuli explained ~7% less of the variance than for upright stimuli. As with the upright stimuli, the linear model showed that thresholds generally increased as the distance from frontal increased (slope: *t*(188) = 14.38, unstandardized *β* = 0.27, *p* < .001, 95% CI = [0.22, 0.30]).

All of the models were very similar with no clear winner (see Figure 6.4a and Table 6.2). In general, adding the M regressor improved the explained variance, but this did not necessarily lead to a significant improvement or a more parsimonious model (as indicated from the AIC and BIC scores). There was also a negligible difference between the *M* + *linear* and *M* + *cosine* model.

6.4.3 Inversion effect

From visualising the data, and from the regression models, it appears that the inversion effect increased more when further from frontal. This can be seen in beta coefficients of the slopes for upright (unstandardized *β* = 0.15) and inverted (unstandardized *β* = 0.27).

To begin to unpack the interaction, an inversion effect for each view was calculated by subtracting the thresholds for inverted stimuli from those for upright stimuli. Linear mixed models were again constructed using linear angle, cosine of the angle, and inter-feature distances to predict the inversion effect.

The linear model showed that as the distance from frontal increased the inversion effect increased (*t*(188) = -5.63, unstandardised *β* = -0.11, *p* < .001, 95% CI = [-0.15, -0.07]).

The inversion effects were best explained by the inter-feature distance and linear models. The inter-feature distance model explained slightly more of the variance (cond. $R^2 = 0.515$, cond. $R^2_{\text{adjusted}} = 0.505$) than the linear model (cond. $R^2 = 0.503$, cond. $R^2_{\text{adjusted}} = 0.500$) but was not significantly better in a SLRT. The AIC and BIC were lower in the linear models. Adding the M regressor decreased the model fit. In summary, the inversion effect could be generally explained by an approximately linear model as the angle from frontal increased.

Interestingly, while the magnitude of the inversion effect seemed to increase with the angular distance from frontal, the effect was proportional to the threshold for upright stimuli. The threshold for inverted stimuli was on average 1.48*x* the threshold for upright stimuli, regardless of viewpoint (see Figure 6.4c). This was confirmed using another linear mixed model, again with a fixed slope and random intercepts, and the simple scaling factor (inverted threshold ÷ upright threshold) as the dependent variable. The simple scaling factor did not significantly vary as a function of viewpoint (slope = $t(188) = -1.83$, $p = .069$, unstandardized $\beta = -5.48\text{x}10^{-3}$, 95% CI [$-0.04\text{x}10^{-2}$, $1.13\text{x}10^{-2}$]).

Allowing the slope to vary across participants did not significantly improve the model ($\chi^2(2) = 2.00$, $p = .368$). In contrast, fixing the intercept significantly decreased the model fit, ($\chi^2(1) = 87.89$, $p < .001$), increased Akaike's Information Criterion from 10.84 to 96.73, and decreased conditional R^2_{adjusted} from 45.78% to 0.42%. Overall, this shows that the relative magnitude of the inversion effect did not differ significantly across views but did by participant.

6.5 Discussion

In this experiment we sought to test if there was a systematic difference across viewpoints in the thresholds for discriminating veridical from warped versions of the same image of a face. These thresholds would reflect which views are best represented in the brain. Based on the prior work (De Souza et al., 2005; Marotta et al., 2002; O'Toole et al., 1998; Perrett et al., 1991; Troje & Bülthoff, 1996; Van der Linde & Watson, 2010) we hypothesised that thresholds would be smaller for frontal, ¾ and profile than intermediate views. Furthermore, we hypothesised that thresholds would be higher for inverted faces and that the inversion effect would be larger for intermediate views due to view interpolation being less efficient.

Overall, we found that there was a general increase in the threshold as the angular distance increased from frontal and that the threshold was higher for inverted faces than upright faces. Contrary to our hypothesis we also found some indication that thresholds were higher for frontal, profile, and ~3/4 (40^o) stimuli than intermediate views. We did not find evidence to suggest that the inversion effect was stronger for intermediate views.

The general increase in threshold for non-frontal faces is congruent with prior evidence that there are more neurons tuned to frontal than profile views (Dubois et al., 2015; Perrett et al., 1991) in view-dependent regions, allowing for a more efficient representation (although see Hasselmo et al., 1989). It is also consistent with evidence that frontal and ¾ faces are more easily detected in natural scenes compared to profile (Burton & Bindemann, 2009). Burton and Bindemann's findings suggest that our results are not due to symmetry or both eyes being present as their effects were observed even when half of the image was occluded.

The lower thresholds for more frontal over more profile faces is also consistent with the amount of exposure we have in the natural world to different viewpoints. Oruc and colleagues (2019) reported that we are exposed to more frontal views than any other view, when calculated per degree. However, under their criteria 'frontal' faces only occupy a very narrow range (±10°). When categorised based on the visibility of the eyes and ears, the ¾ view is the most prevalent category, and profile the least. The slight decrease in threshold that we observed for stimuli at 15-30° may also therefore be explained by exposure, as this range is near frontal, yet would be categorised as '¾' according to the conditions used by Oruc and colleagues. The increase in threshold at ~40° however is not well explained.

The evidence for a behavioural benefit for slightly non-frontal faces is limited, however there is evidence that landmark detection and face identification in DNNs is optimal for slightly non-frontal (10-20°) views (Choithwani et al., 2023). There is also evidence that in both humans and DNNs lipreading is easier at ~30^{o} (Lan et al., 2012). All three of these scenarios may explicitly benefit from the processing of 3D information. For instance, landmarking may be easier with access to the additional depth information from a non-frontal image, although the optimal view was not too far from frontal. Lipreading may also be easier at 30^{o} due to having good visibility of the full width of the mouth as well as additional depth information about lip protrusions. In our study it may have been easier to distinguish warped facial features from veridical shapes using the additional structural information available at 15-30^{o} from frontal.

As well as a slight decrease in threshold for ~15-30^{o} stimuli relative to frontal and ¾, we also observed a slight decrease at ~50-60^{o} relative to ¾ and profile. This decrease is not well explained by exposure but perhaps the advantage for both of these intermediate views could be due to overlap in broad tuning curves (e.g., Perrett et al., 1991). We hypothesised that the viewpoints that neurons would be preferentially tuned to (based on macaque studies) would be the viewpoints with the lowest threshold, but it is possible that the overlap between broad tuning curves for intermediate views allows them to be more efficiently represented.

The presence of lower thresholds at intermediate views does not refute the neurophysiological evidence of neurons preferentially tuned to frontal, 45^{o} and profile (De Souza et al., 2005; Perrett et al., 1991) in favour of intermediate views, but it may indicate that more neurons are tuned to these intermediate views than previously sampled. Neurons responding maximally to these intermediate views alongside the overlap in broad tuning curves for frontal, ¾ and profile views may result in lower thresholds.

The thresholds being higher for inverted faces than upright faces is consistent with a wide range of evidence (Eng et al., 2017; Garrido et al., 2008; McCleery et al., 2008; Thornton et al., 2011; P-L. Yang & Beck, 2021, 2023; R. K. Yin, 1969) and shows that our task tapped into higher-level face processing mechanisms. Moreover, it supports the prior evidence that the perceptual representation (before any cognitive judgement about identity) is optimised for upright faces. This includes the work by Yang and Beck (2021, 2023) who also found larger thresholds for inverted faces and from whom the methods of this current study were based. The fact that we too saw a substantial inversion effect is reassuring given that we manipulated warp level rather than stimulus duration and had many more conditions. Had we not seen an effect of inversion then the lack of clear systematic effects of viewpoint (other than the general increase with view away from frontal) would likely result from a dependence of lower-level visual features rather than higher-level features. It is worth noting however that the manipulation of the warp level may have made the task more cognitive. The task did not require participants to identify the face, and it should have required more image-based assessments, but it may have required more cognitive judgements about the natural shape of the actor’s face.

The magnitude of the inversion effect increased as distance from frontal increased, but the threshold for inverted stimuli was proportional to those for upright stimuli. This suggests a consistent, proportional disadvantage for inverted compared to upright faces. It would be interesting to see if this relative detrimental effect is seen in future studies. Similar results were observed by Favelle and Palmisano (2012) who in Fig 3 present data showing that the sensitivity for matching inverted faces is approximately proportional to the sensitivity for upright faces (upright = ~1.4x inverted). The relative difference is slightly larger for frontal faces (~1.6x), but this may be because the first stimulus in each stimulus pair was frontal. In contrast however, Watson and colleagues (2005) found a greater degree of variability in the effect of inversion across view.

We did not find any additional detrimental effect on the inverted stimuli for views that were either classed as intermediate *a priori* or the prototypical views (0^{o}, 45^{o}, 90^{o}). If the interpolation account (H. H. Bülthoff & Edelman, 1992) of face processing is accurate then, according to our data within a face detection paradigm, interpolation is as efficient for inverted stimuli as it is upright, once the general inversion effect is accounted for. This is potentially consistent with evidence that identity discrimination is equally affected in changes in view for both upright and inverted faces (Wright & Barton, 2008). In contrast, other research with unfamiliar faces has shown that viewpoint generalisation is worse for inverted faces (Moses et al., 1996). In the current experiment, most participants were familiar with the actor and thus more robust interpolation for inverted faces might be possible. The current task was also based on a more perceptual task (detection) rather than a more cognitive task (recognition) for which interpolation may be more efficient. That said, our results for inverted stimuli were clearly noisier than for upright stimuli, so further work would be needed to form any conclusions on this matter.

There are a few confounds that may contribute to an increased threshold at ~40°, however none provide a perfect explanation for the findings. While 20 different sets of warps were created, it is possible that warping had less of an effect on the features or contours necessary for discrimination at ~40° compared to other views. However, the internal features are closer to the external contours than for more frontal faces, but not as close as for more profile faces, so it is not clear why this configuration would be particularly affected. Participants may have also changed response strategy at ~40°. For frontal faces, the internal features of the two face stimuli are approximately equidistant from fixation, so it is easier to attend to both faces concurrently. At profile, the internal features of one face are closer to fixation than the other, meaning that participants may have attended to the closer face and decided if that was morphed or veridical. At 40°, participants may have been torn by which strategy to use as neither were optimal, increasing the threshold. The effects of warping and changing strategies, however, would be expected to be more gradual and should have affected both upright and inverted faces similarly. While a slight peak is present at ~40° for inverted faces this is less clear than for upright faces. A spatial jitter was also introduced which would have further reduced sudden effects due to changes in strategy. One might also expect a decrease in thresholds if participants were able to use both strategies concurrently rather than just one.

A direction for future work would be to compare the thresholds across views for familiar and unfamiliar participants. Depending on how view-invariance is achieved, one may find different patterns in the thresholds for familiar and unfamiliar participants. Following the results from Yang and Beck (2021, 2023), one might expect that any benefit for specific viewpoints might be stronger for familiar over unfamiliar faces. On the other hand, weaker effects across viewpoint might be observed for familiar observers than unfamiliar due to having a better view-invariant representation.

In summary, we found a clear increase in the thresholds for inverted faces and faces closer to profile compared to upright and closer to frontal. Contrary to our predictions, slightly higher thresholds were observed for frontal, ¾ and profile views relative to intermediate views. More research is required to see if this finding replicates. If it does, this may provide useful insight into which viewpoints are preferentially represented and the efficiency of view-interpolation.

Chapter 7 General discussion

Recent research has shown strong evidence for norm-based coding in macaques through axis coding (Chang & Tsao, 2017; Koyano et al., 2021), allowing faces to processed in a linear face space. This doctoral research assessed two aspects of linear face space. Firstly, Chapter 3 assessed what happens when face space is taken to the extreme, by presenting participants with hyper-caricatured faces. This fMRI experiment showed that caricatured faces beyond the realm of natural plausibility activate cortex typically associated with object processing. The second stream of work addressed how to incorporate view-invariance into a 2D face space, specifically in the domain of facial motion. In Chapter 4, various models that tried to reconstruct facial motion across viewpoints were described, concluding with a two-step model capable of accurate reconstructions whilst adhering to some constraints of biological plausibility. Chapter 5 then discussed how mirror-views might be represented within a view-invariant face space, expanding briefly on the structure of the models in the previous chapter. Chapter 6 began to test which viewpoints might be preferentially represented in the human brain, to help guide future construction for models of view-invariance such as those described in Chapter 4.

7.1 The effects of caricaturing in face space

In the fMRI study assessing the response to hyper-caricatured faces, we anticipated that as the caricature level increased, the BOLD response in the FFA and other face selective areas would increase.

This was based on single unit recordings in macaques (Chang & Tsao, 2017; Leopold et al., 2006) although there was mixed evidence for the shape of the response function in humans (Carlin & Kriegeskorte, 2017; Loffler et al., 2005; McKone et al., 2014; Susilo, McKone, & Edwards, 2010).

A small increase in the BOLD response for highly caricatured faces over more average faces was observed in the right FFA, but this was not significant when considering all caricature levels and only upright stimuli. The response pattern, however, increased in consistency in the right FFA as caricature level increased, perhaps through enhancing the signal-to-noise ratio. In comparison, there was an increase in both response amplitude and response pattern consistency in object-selective cortex.

Although the consistent response amplitude in face-selective areas to different caricature levels overall was not predicted based on a norm-based representation of face space, it is not clearly congruent with an exemplar-based model either (Lewis, 2004; Valentine, 1991). An exemplar-based model would predict the highest response for the most average faces, as these would fall into the region of face space with the highest density and therefore the most overlap with known exemplars, and the smallest response for highly caricatured faces. While we did not see a clear increase in response amplitude with increasing caricature level, neither did we see a decrease. If faces are processed with completely separate spaces, say with ‘islands’ for familiar identities (Hancock, 2021) that are not situated in an exemplar code, then we perhaps would predict a consistent response across all caricature levels for these unfamiliar stimuli, as all caricature levels would sit within the sea of unknown faces. However, we can generally distinguish different identities and different caricature levels, so there must be some common space in which all faces are processed to at least some degree.

Within a norm-based representation, we have already described how the relatively consistent response amplitude to different caricature levels may have occurred through a balance of some cells increasing and others decreasing in firing rate. However, there may also be a balance between the initial response, and later inhibition to more average stimuli. In macaques, Koyano and colleagues (2021) found initial ramp tuning to stimuli at a greater distance from the average, with cells either firing more or less relative to average depending on whether the stimulus was on the preferred direction of the neurons axis. At a later timepoint they then found a downregulation of the response to more average stimuli. Our results may therefore reflect a balance between the initial responses and the downregulation. Further studies using MEG or EEG could explore this possibility as it would help to elucidate the neural coding underlying human face processing.

Alternatively, it is possible that the FFA does not code faces within a face space representation. Baseler and colleagues (2016) suggested that the FFA plays an intermediate role in alignment prior to further processing (perhaps in a face space manner) in more anterior areas. Different caricature levels may therefore require different neural populations for the alignment, leading to no net effect of caricature level. It is possible that the distortions seen during electrical brain stimulation of the right FFA (Parvizi et al., 2012; Rangarajan et al., 2014) reflect errors in alignment rather than accentuating particular dimensions of face space as suggested in the introduction. The results and discussions here are, however, incongruent with the recordings in macaque area ML (Chang & Tsao, 2017), thought to be the homologue of human FFA (Tsao et al., 2003), and the already mentioned results from (Loffler et al., 2005). Yet they are consistent with evidence of an intact face space representation in a patient with prosopagnosia who had damage to the right fusiform gyrus (Rivest et al., 2009). We cannot comment on face representations in more anterior areas as our slice prescription was focused on the OFA, FFA and pSTS so the anterior temporal cortex and IFG were outside the field of view.

Despite generally consistent responses across caricature levels in face-selective regions, we did observe some trends worth highlighting. Firstly, as already highlighted in Chapter 3, we observed evidence that the response in the right FFA increased to highly caricatured faces whereas the left FFA remained unperturbed. This was shown in the ANOVA comparing upright and inverted faces at the highest and lowest caricature levels. A follow-up comparison showed that the difference in hemispheres was still significant when only considering the upright stimuli (*t*(8) = 2.67, *p* = .029) prior to correction.

The left FFA's indifference to caricature level is potentially consistent with functional differences between the left and the right FFA as shown by electrical stimulation (Rangarajan et al., 2014) and fMRI (Rossion et al., 2000).

From visually inspecting Figure 3.4, there also appeared to be an increase in the response amplitude for caricatures on the naturalness boundary in the right FFA. A subsequent analysis revealed a significantly larger response in the right FFA than left to these stimuli ($t(8) = 3.82$, $p = .005$). This is potentially compatible with the peak in aftereffect strength around the naturalness boundary seen in prior work (McKone et al., 2014) and may further demonstrate functional differences between hemispheres (Rangarajan et al., 2014; Rossion et al., 2000). However, this trend was also present in object and scene-selective regions, potentially reflecting a global enhancement in visual attention; participants may have been consciously aware that these stimuli closely matched their perceptual boundary from the behavioural study. This is compatible with general evidence for right hemisphere dominance in attentional processes (e.g., Heilman & Abell, 1980) although is less consistent with evidence that the response in FFA increases bilaterally when attention is guided to faces (Wojciulik et al., 1998). It is also not consistent with evidence showing that attention to, and mental imagery for, faces over houses selectively activates face-selective cortex (O'Craven et al., 1999; O'Craven & Kanwisher, 2000; Wojciulik et al., 1998) and conversely that the parahippocampal place area is selectively activated by attention to houses (O’Craven et al., 1999; O’Craven & Kanwisher, 2000).

Participants may have attended to more local, lower-level visual features to discriminate plausible from implausible faces. This would likely have less of a specific effect on face-selective regions if the features attended were lower-level visual properties. However, if they were attending to more local *facial* features then this would instead predict an increase in response in the left FFA, which is more sensitive to part-based information (Rossion et al., 2000).

The lack of hemispheric differences for mildly caricatured faces (+1SD and +3SD) further muddies the argument about functional differences between the left and right FFA. Why would highly caricatured faces and faces on the naturalness boundary selectively enhance the response in the right FFA but not the left? And why do intermediate caricature levels not elicit a difference between hemispheres? As already noted, changes in visual attention do not explain these hemispheric differences well (Wojciulik et al., 1998). Moreover, if the response to highly caricatured faces in the right FFA indicated a difference in face space representations then the BOLD response would be expected to change in a more linear fashion across caricature levels.

Although the responses to faces on the naturalness boundary (0SD) and highly caricatured (+6SD) faces were larger in the right than left FFA, the responses in the right FFA to these levels were not significantly different from the other caricature levels once corrected for multiple comparisons (all *p* > .034). Further studies would therefore be needed to explore whether, and if so why, the right FFA shows an increase in response to caricatures on the boundary of natural plausibility and highly caricatured stimuli but not intermediate caricature levels.

While some differences across hemispheres were present for certain caricature levels, there was not a main effect of hemisphere overall. Studies have often shown that the FFA is more consistently found in the right hemisphere than the left (Kanwisher et al., 1997; Pitcher et al., 2011, 2023; Sliwinska et al., 2020) and is larger in the right hemisphere (Bukowski et al., 2013; Pitcher et al., 2023) supporting a right hemisphere dominance in face processing. In contrast, we did not see a difference in volume between the left and right FFA, nor did we see a main effect of hemisphere overall in the response amplitude or, in a follow-up analysis, in the response pattern consistency. However, not all prior studies show right hemisphere dominance (e.g., Fox et al., 2009; Grill-Spector et al., 2004; Ishai et al., 2002) and there is evidence that a right hemisphere advantage is only seen for certain tasks (Ellis & Young, 1983). The lack of a right hemisphere advantage overall was therefore not of concern in our study, and possibly reflected the task not being directed to the faces.

In contrast to the response amplitudes, which did not differ significantly across caricature level within the FFA or face-selective regions generally, we did observe an increase in the response pattern consistency in the right FFA. This suggests the right FFA is sensitive to caricature level, and that increasing the caricature level might elicit a more consistent response profile. This may therefore suggest the presence of a face space representation in the right FFA but we cannot discount the possibility that the FFA is involved in alignment procedures (Baseler et al., 2016).

In object-selective cortex, both a significant increase in the response amplitude and the response pattern consistency was observed as the caricature level increased, particularly for hyper-caricatured stimuli. This may be due to low-level properties such as certain shapes or colours, or it may be that object-selective cortex provides some functional benefit to recognising caricatured faces.

It has long been debated whether higher-level visual processing is modular or distributed (Haxby et al., 2001) so it seems possible that caricatures might exacerbate the response to faces in cortex typically defined as object-selective, potentially through similarities of lower-level features. As faces have been argued to sit within a more general object-space (Bao et al., 2020), regions outside of face clusters respond to faces (Haxby et al., 2001) and the OFA responds to face-like objects (Decramer et al., 2021), this seems plausible. Indeed, the response to faces compared to baseline in the localiser scan of our study

(Supplementary Figure 3.6) shows that many object-selective voxels also respond to faces. These object-selective regions may even be recruited for cross-communication with face-selective voxels for trying to process these ‘unusual’ faces. Disrupting the face processing network by applying transcranial magnetic stimulation (TMS, Barker et al., 1985) to the OFA reduces the response to both faces and objects in the object-selective LO (lateral occipital, Pitcher et al., 2014) suggesting that face and object-selective regions are connected and overlap in function. That said, evidence from electrical stimulation in macaques suggests that the communication between neurons is highly selective to face-selective regions, with little cross communication (Moeller et al., 2008). If cross-communication is present, however, then this could explain the behavioural benefits of caricaturing discussed in the introduction if both face- and object-selective areas are working in tandem to process these more unusual stimuli.

7.1.1 Directions for future work

As already mentioned, some future work could look to expand this study through the inclusion of M/EEG. The precise temporal resolution would allow these studies to investigate the presence of initial ramp responses to caricatured faces followed by lateral inhibition to average faces (Koyano et al., 2021). This would help distinguish norm-based and exemplar representations of face space.

Like previous research (Carlin & Kriegeskorte, 2017; Loffler et al., 2005) we used unfamiliar faces, however it would also be beneficial to assess the effect of caricaturing on familiar individuals. Familiar faces increase the response in the OFA and FFA compared to unfamiliar faces (Eger et al., 2005; Ewbank & Andrews, 2008; although see Minnebusch et al., 2009) and familiarity biases the magnitudes of perceived distinctiveness, attractiveness, likeability and trustworthiness compared to their respective anti-faces (Faerber et al., 2016). In contrast, unfamiliar faces show equivalent ratings across face/anti-face pairs (Faerber et al., 2016). Familiar faces might therefore elicit a sharper rise in response amplitude, perhaps revealing a more substantial effect of caricaturing than found here. Caricatured familiar faces might also provide evidence for separate spaces or subspaces capturing within-identity variation (Burton et al., 2016; Hancock, 2021). This would be particularly insightful if caricaturing is coded relative to identity-specific norms rather than a norm of all encountered faces.

Further research should also investigate why hyper-caricatured faces activate regions typically associated with object processing. If the object-selective cortex is functionally involved in face processing, then caricaturing may provide a route to helping those with prosopagnosia. One would of course need to establish whether the object-selective cortex does contribute, and if so, which regions specifically, perhaps using non-invasive brain stimulation techniques such as TMS. In the current study all object-selective voxels were grouped into one ROI. Recognition for hyper-caricatures could also be assessed in individuals with prosopagnosia.

7.1.2 Is caricaturing a route to helping those with prosopagnosia and other conditions?

If object-selective cortex poses a functional benefit to recognising highly caricatured faces, then hyper-caricaturing may benefit those whose face recognition streams have been damaged or function differently, as is often the case in acquired prosopagnosia (AP) and developmental prosopagnosia (DP) respectively.

If caricaturing is beneficial for individuals with prosopagnosia, then similar portable technologies to those mentioned in Chapter 4 might be beneficial. Chapter 4 mentioned that wearable technologies will soon be able to help convert visual information into auditory speech, to help those with hearing difficulties. Similar technologies may help individuals with face recognition difficulties. Dawel and colleagues (2019) suggested that caricatures may help improve recognition in patients with prosopagnosia, as well as individuals with vision loss, such as central vision loss in age-related macular degeneration. They also described how technology could caricature faces in real-time, such as on a phone or tablet, or with augmented reality glasses. As noted by Dawel and colleagues, this technology would need to work under any viewing conditions including across large changes in pose, but there is currently no software effective enough at this, in part due to the necessity for fast and accurate landmark detection. These are the same requirements needed by, and the same limitations of, software for transforming facial motion across viewpoint and into auditory speech. The technologies could even be combined; caricaturing could be applied to facial motion to improve lip-reading and similarly, the multi-view PCA model could be used for frontalisation before caricaturing information diagnostic of identity.

The benefit of caricaturing, and therefore the suggested technologies, firstly depend on whether an intact face space representation is present. Individuals with DP generally show an intact face space representation (Leib et al., 2012; Nishimura et al., 2010; Robson et al., 2018; Susilo, McKone, Dennett, et al., 2010 b), although some research suggests a modified face space (Palermo et al., 2011; Robson et al., 2018). In contrast, the evidence for intact face space representations in patients with AP is mixed (Nishimura et al., 2010; Rivest et al., 2009). This distinction may influence how caricatures are used, if at all. Moreover, the amount of caricaturing required might also vary. For instance, if no face space representation is present, but object-selective cortex provides some functional benefit for recognising highly caricatured faces, then one possibility would be to hyper-caricature faces and force the processing to occur in object-selective cortex. In contrast, if a face space representation is present, then less caricaturing might be necessary.

In general, a lack of a face space representation in acquired prosopagnosia is not surprising (e.g., Nishimura et al., 2010). Face space representations should reside within face-selective regions of cortex, which are often lesioned in AP. However, evidence for an intact face space has been seen in some cases (e.g., Rivest et al., 2009). The lesion in this patient included the right fusiform gyrus suggesting that if a face space representation is present, it may not reside in the right FFA. Yet, another patient with damage to the right FFA (Behrmann & Williams, 2007) showed no indication of an intact face space (Nishimura et al., 2010). Both patients had damage extending more anterior to the FFA, as well as to other areas, so it is not clear where face space is represented.

Caricaturing has been shown to benefit healthy controls (Itz et al., 2017; although see Minnebusch et al., 2007), individuals with poor face recognition, but not necessarily prosopagnosia (Limbach et al., 2022; Powell et al., 2008) and individuals with age-related macular degeneration (Lane et al., 2018). Caricaturing also provides larger benefit for older adults than younger adults despite worse recognition, suggesting that face space may be present but distorted in older adults (Dawel et al., 2019) and that older adults and those with worse recognition would benefit most from caricaturing technologies. Despite this, and despite evidence for intact face space representations, at least in DP, there is mixed evidence whether caricaturing benefits face processing in prosopagnosia.

Three-dimensional hyper-caricatures, containing both shape and texture distortions, have been found to improve behavioural performance and increase the N170 in patients with DP (Minnebusch et al., 2007) and they elicit comparable BOLD responses to controls, despite DPs showing more variable responses to veridical famous and unfamiliar faces (Minnebusch et al., 2009). This indicates the representation of caricatures is more stable in DPs and potentially corroborates our findings of more consistent response patterns in the FFA to caricatured faces. In contrast to this, and in contrast to the studies discussed previously, controls showed some behavioural deficits for caricatures compared to veridical pictures, alongside weaker N170s (Minnebusch et al., 2007). Controls also showed larger BOLD responses to caricatures and unfamiliar faces than famous faces (Minnebusch et al., 2009) which the authors interpreted as increased processing costs of viewing novel (caricatured and unfamiliar) stimuli.

Other research, however, suggests that caricaturing does not help those with prosopagnosia. Some individuals with prosopagnosia are able to recognise cartoons, both from silhouettes (M. Cook et al., 2019) and line drawings (Rivest et al., 2009) yet are unable to recognise line-drawn caricatures (Rivest et al., 2009). Others are not able to recognise faces from line-drawn caricatures or cartoons (de Gelder & Rouw, 2000). Sergent and Signoret also reported that individuals with AP were worse at recognising caricatures than controls, and patients reported that they “found the tasks with caricatures very frustrating”, and caricatures “made even less sense to them than normal faces”, and “were more difficult to recognise” (Sergent & Signoret, 1992, p. 384). Matching caricatures to their veridical face was also impaired, suggesting that their face space had been disrupted. The line-drawn caricatures used in these studies, however, are limited or even devoid of texture, which appears to be important for recognition and learning (Itz et al., 2017; Limbach et al., 2022), so they may not engage the same processes as more image-realistic caricatures made by computationally extrapolating face space.

Instead of caricaturing, Powell et al (2008) found that only directing attention to distinctive features improved performance for an individual with pure prosopagnosia (W.J.). This is despite observing benefits of caricaturing for brain damaged patients who show face processing deficits, but who do not strictly have prosopagnosia. Combining caricaturing and part-based attention might improve performance further, although W.J. was almost at ceiling with part-based attention in this particular task. Guiding attention away from the mouth and to the eyes has also been found to improve face matching in AP (Ramon & Rossion, 2010), suggesting that some of the deficits stem from over-attention to the mouth. If combined with caricaturing however, the focus of directed attention may of course differ depending on what features are particularly distinctive.

Overall, there is mixed evidence that caricaturing helps those with prosopagnosia, possibly due to the methods and particular stimuli used. More realistic textured stimuli seem to suggest that caricaturing faces can help those with prosopagnosia (Minnebusch et al., 2007, 2009), and caricaturing texture has also been shown to improve face processing (Itz et al., 2017; Limbach et al., 2022). Therefore, if caricatures are to help those with prosopagnosia, retaining and caricaturing texture is likely crucial.

There is limited evidence whether caricatured faces recruit cortex outside of typical face-selective regions. But, some evidence has shown that caricaturing the distances between features during training leads to better recognition alongside increased connectivity between face- and non-face-selective regions in DP (DeGutis et al., 2007).

Even if the object-selective cortex does not functionally contribute to caricature processing *per se*, the intentional caricaturing of stimuli to activate object-cortex might prove useful through plasticity. If caricatured faces activate object-selective cortex, then over time this cortex may learn to more effectively represent these stimuli, not necessarily as faces *per se* but as a class of object. It may then even be possible to decrease the amount of caricaturing required. In other words, caricatures might be able to guide object-selective cortex into processing faces. Current evidence for this as a possibility is mixed. Hadjikhani and de Gelder (2002) studied responses in three individuals with prosopagnosia from either birth or early childhood and found no evidence of cortex selectively responding to faces over objects. Yet, for one individual the (typically object-selective) region LO responded equally to faces and objects compared to houses, suggesting that LO had assimilated some ability to process faces, though not enough to sufficiently enhance face recognition.

The suggestion of encouraging faces to be processed in object-selective regions is further complicated as face and object processing deficits are often co-morbid (e.g., Barton et al., 2003; Behrmann &

Williams, 2007; Farah et al., 2000; Svart & Starrfelt, 2022). Also, even in DPs there is substantial heterogeneity in the responses to faces within the OFA and FFA (Minnebusch et al., 2009) before even considering object-selective regions.

In summary, we found evidence that object-selective cortex responds to hyper-caricatured faces, and thus posed the question of whether hyper-caricatured faces might help those with prosopagnosia through recruiting object-selective cortex. Patients with developmental prosopagnosia often show evidence of intact face space representations, and caricaturing has sometimes proven useful in improving recognition both in DP and AP, showing that caricaturing can be beneficial. Furthermore, the evidence for increased functional connectivity between face- and non-face-selective regions with training of caricatured configurations, and the assimilation of face responses in object cortex in prosopagnosia, indicates that engaging object-selective regions for processing caricatures might be possible. The efficacy of such a strategy would of course depend on the presence of damage to object-selective regions or object processing deficits, so would be highly individualised. Nevertheless, this provides an exciting and highly beneficial line of research.

7.2 Learning view-invariant representations in face space: the multi-view PCA models

Moving away from caricaturing, chapters 4 to 6 assessed view-invariance and facial motion processing in face space. In Chapter 4 the aim was to build a biologically plausible model that could reconstruct facial motion across changes in viewpoint, based on the premise of using 2D, view-specific representations rather than a 3D representation. The aim of the work was to expand on and accumulate methods used by Beridze (2021), Lan et al (2012) and Lucey et al (2007), Scholes et al (2020) and Burton et al (2016). Model 5, the two-step model, was the most promising of those described and was able to reconstruct motion well across all 5 viewpoints whilst adhering to biological constraints, primarily that the model was never exposed to all viewpoints simultaneously.

The first layer of the two-step model contained view-specific spaces, however rather than directly projecting frames across spaces (Beridze, 2021) the associations between the spaces were learned, akin to in Lan et al (2012) and Lucey et al (2007). However, it was not possible to learn these associations directly from the components due to using the full face rather than just the mouth, as the components were less spatially aligned across views. Instead, the relationships between the loadings on the components in neighbouring spaces were learned. The model also improved biological plausibility over Lan et al (2012) by only learning the associations between neighbouring views and not distal views.

Although cross-view reconstructions could be made in the first layer, it was in essence a view-dependent layer, as no single unit (e.g., component) contained knowledge of all 5 viewpoints. This layer is commensurable with evidence of view-dependent neurons in posterior regions of the macaque STS (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991) and in more posterior regions of the human face processing system (Ewbank & Andrews, 2008; Fang et al., 2007; Guntupalli et al., 2017; Natu et al., 2010; Pourtois et al., 2005).

The second layer of the model then forms the view-invariant layer, with each unit possessing information about multiple viewpoints. It is possible to reconstruct motion well across views with this model, which provides a potential mechanism for how view-invariance might be achieved in the brain. The view-invariance in this layer is commensurable with evidence of view-independent neurons in macaques (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991) and evidence of view-invariance in more anterior regions of the human face processing system (Anzellotti et al., 2014; Guntupalli et al., 2017; Pourtois et al., 2005).

For example, Chang and Tsao (2017) found that neurons in the more anterior region AM showed axis coding to multiple viewpoints. The second, multi-view layer of Model 5 shows how axis coding can occur for multiple viewpoints from a set of 2D representations, in essence by pooling the responses from the view-dependent layer. The representation did not need to form a 3D model to show corresponding changes in appearance across the views within a given axis.

The two-step model is also compatible with adaptation aftereffects. The first, view-dependent layer would give rise to view-selective adaptation aftereffects (J. Chen et al., 2010; Jeffery et al., 2006) particularly for unfamiliar faces (Jiang et al., 2007) and would allow the specific, contingent aftereffects observed by Welling and colleagues (2009). It is also compatible with evidence for multi-channel coding of viewpoints (Fang & He, 2005). The learned associations between neighbouring spaces can then help explain why aftereffects translate across small changes in view (Jiang et al., 2007).

Once the multi-view representation in layer 2 is developed then greater transfer of the adaptation aftereffects across views can occur, if the adaptation effects either propagate through to, or occur in, the view-independent, multi-view layer. For this to occur, there would need to be sufficient experience with the person to learn the associations between the separate view spaces and build the multi-view representation. This is supported by evidence showing that increased familiarity, even with static stimuli, can increase the transfer of aftereffects across views (Jiang et al., 2007), although they did not test for adaptation effects across large rotations. The effect of familiarity on adaptation aftereffects across large rotations has yet to be thoroughly explored.

The amount of transfer across views would of course depend on the aftereffect being explored and whether separate spaces are required for each familiar individual. If there is one identity-general space, then the effect of familiarity would be less than if identity-specific spaces were necessary.

The multi-view face space model has been constructed based on the assumption of opponent, norm-based coding around an average, neutral expression. However, the effects of caricaturing facial motion shown by Furl and colleagues (2020) suggest that facial actions are also coded around action-specific norms, supporting instead exemplar coding. This is also supported by prior behavioural evidence showing that caricaturing motion relative to a static norm results in undesired changes in the perceived expression, whereas caricaturing motion around a dynamic expression-specific average selectively increases the perceived intensity of that expression (H. Hill et al., 2005). However, if different actions are processed in a common space using exemplar coding rather than completely separate spaces, then the PCA spaces constructed could provide the set of common dimensions for exemplar coding to occur in. The average trajectories of facial motion could be learned within this space to form the local, utterance-specific norms.

7.2.1 Future directions

While the multi-view PCA models, namely Models 3 and 5, can reconstruct motion well across views, there are many areas that can be expanded.

Firstly, the videos were captured under constrained viewing conditions and the actors remained fairly neutral in expression throughout. To make the model more effective in the real-world, videos would need to be captured under more varied viewing conditions and demonstrate a range of expressions, alongside variations in appearance such as health, makeup, and facial hair.

Secondly, one could explore whether one identity-general model would be better than separate identity-specific models. Identity-specific models have the benefit of capturing important idiosyncratic information (Burton et al., 2016) useful for identity, emotion and speech processing (H. Hill & Johnston, 2001; Rosenblum et al., 2002, 2006; Sheffert & Olson, 2004). They are also more commensurable with suggestions that familiar faces are represented as discrete islands (Hancock, 2021). Facial motion of a novel person could be processed in one or more familiar spaces if they share resemblance (Hancock, 2021). On the other hand, separate spaces have additional storage requirements over a single identity-general model. They also raise the question of how completely novel faces are processed if bearing no resemblance to any known face.

Thirdly, the models should be expanded to work incrementally and be updated with new data. In the current models the videos were provided as a single batch, whereas real visual input would appear incrementally. A MATLAB implementation of incremental PCA (Wai, 2021), which takes a small batch of initial frames, performs PCA and then updates the PCA space with new information, showed promising results in pilot testing, but this was only tested briefly for Model 3.

Example components and reconstructions can be seen in the supplementary videos. It would also need testing for Model 5 (the two-step model) as one would need to allow both layers to be updated, as well as the relationships between the separate spaces. If there is also an identity-general space that uses an average template as a single prototype for guiding recognition (Abudarham et al., 2019; Jenkins & Burton, 2011) then as the models update, the average representation for a given identity can also be fed back to improve their prototype and improve cursory recognition from the identity-general space (Burton et al., 2005).

The models would also need to be tested with intermediate views. While there was some rigid head motion and rotations seen by each view, the models would need to be able to reconstruct motion from, and possibly in, intermediate views such as 12°. One could, for example, directly project frames for these viewpoints into the spaces after subtracting a mean of that view (12°) or the mean of the space projected into (e.g., 0° or 22.5°). If motion cannot be sufficiently represented in these spaces, then the model would not necessarily provide a good account of viewpoint interpolation.

As well as expanding the model, there are also some assumptions that need to be tested. In these models, it was assumed that facial motion is processed in a view-invariant manner. This was in part assumed because motion helps achieve view-invariance (O'Toole et al., 2002; T. Watson et al., 2005), as was outlined in the introduction.

This could be tested by establishing whether motion in one viewpoint primes the representation of the action in another. If it does, then this suggests that motion is processed in a view-invariant manner.

Further work should also expand on the experiment in Chapter 6 to further investigate the presence of prototypical viewpoints, to guide which views are used in the model. Ideally, similar experiments would be performed with different actors, a larger sample and a wider range of stimuli and tasks.

In this work we used pixel-based information to calculate the PCA spaces to provide an interpretable description of the methods. The methods however should be transferable to vectors containing the output of filters such as Gabor filters, Hermite functions and Gaussian derivatives which have been shown to predict responses in simple cells within cat and primate visual cortex well (Jones & Palmer, 1987; R. Young et al., 2001; R. Young & Lesperance, 2001). The output from these filters can replace the vectors of pixel and warp values. The model should learn how filter outputs are related across views in the same way as for pixel values.

Finally, the current models only contained one side of the face, therefore further improvements could be made by expanding the model to also include the other hemi-view. Some suggestions of how mirror views are included was provided in Chapter 5 and discussions surrounding the expansion of the model to include mirror views are provided in the next subsection.

7.3 Mirror views

As discussed in Chapter 5, evidence of responses to mirror views are taken to suggest an intermediate stage between view-dependence and view-invariance. For instance, neurons more centrally spaced on the posterior-anterior face processing hierarchy seem to respond to mirror views without responding in a fully view-invariant manner (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Meyers et al., 2015; Pinsk et al., 2009). Recent evidence in humans also suggests that mirror views and mirror symmetry are important stages within view-invariant face representations (e.g., Flack et al., 2019; Kietzmann et al., 2017; Rogers & Andrews, 2022). Despite this converging research, it is not yet clear how or why mirror views are used in achieving view-invariance.

How this happens may be highly informative in understanding how faces are represented as there are different ways that mirror views can be used depending on whether the 3D or 2D account of face processing is true. For instance, Chapter 5 considered how mirror views might be represented in a 2D face space, and showed how one hemi-view might be mirror-flipped, so that both hemi-views can be combined into a single space. In Chapter 5 we then briefly outlined how that representation might fit with the multi-view face space models in Chapter 4, particularly in relation to the two-step model. However, the exact representation would need to be elaborated further to understand how the layers in the multi-view space correspond to the 3 apparent stages of face processing: view-dependence, invariance across mirror views, and full view-invariance (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Meyers et al., 2015; Pinsk et al., 2009).

Firstly, it is evident from the neuroimaging results described above and from adaptation studies that mirror views initially need representing separately. Repulsive aftereffects of adapting to a side view changed the perceived direction (left/right) of a near-frontal view to the opposite direction (J. Chen et al., 2010; Fang & He, 2005; Ryu & Chaudhuri, 2006), with adaptor strength peaking at 20-30^{o} (J. Chen et al., 2010). This shows that mirror views would need, at some level, separate representations for repulsive aftereffects across frontal to occur. Moreover, the results show that coding of viewpoint is multichannel at this stage rather than two-pool opponent coding (J. Chen et al., 2010). Interestingly, these repulsive effects are stronger for familiar than unfamiliar faces, but only when the identity remains constant across adaptor and test (Ryu & Chaudhuri, 2006). When identity changed on familiar trials participants were impaired in reporting which way the test stimulus was facing. This indicates separate representations of view for familiar and unfamiliar participants and that the aftereffects for familiar individuals are stronger but identity-specific. This is compatible with having separate identity-specific representations in the first, fully view-dependent layer of the model, but also perhaps indicates a distinct set of spaces for representing unfamiliar faces.

This fully view-dependent representation could be the current first layer of the multi-view model if expanded to include additional separate spaces for the currently unseen hemi-view. It may be the case that half of the spaces (one hemi-view) are processed in one hemisphere, and the other half in the other hemisphere (Corballis & Beale, 2020; Gross et al., 1977). The high level of inter-hemispheric connectivity (Davies-Thompson & Andrews, 2012) would then allow for seamless perception across hemi-views.

If layer 1 of the two-step model is fully view-dependent and has separate spaces for the mirror views, e.g., 9 spaces instead of 5 then another, middle layer would be needed for collating mirror views, but how and to what end? In the combined space model of Chapter 5 we considered the possibility that the final (multi-view) layer of the two-step model may represent a single, combined 'side' of the face, so the additional, middle layer may combine mirror views prior to forming this representation.

The computational benefits of having an intermediate layer would need to outweigh the costs relative to directly constructing the multi-view layer from the fully view-dependent layer. In terms of vectors and matrices, we have shown the potential advantages of collapsing mirror views over representing redundant symmetrical information in separate slots for each hemi-view in the multi-view components. Thus, an intermediate layer could collapse across hemi-views before making the multi-view representation. Each multi-view PCA component is meant to represent a single neuron, so the descriptions of vectors with separate ‘slots’ for separate views and hemi-views make less sense, but the slots can perhaps be thought of as inputs to the neuron. A vector with 9 view-specific slots might represent a neuron with 9 connecting synapses, whereas a condensed vector with 5 slots might represent a neuron with only 5. This is obviously an over-simplification, but the point is that the benefit of an intermediate stage might be through reducing the number of synaptic connections to the view-invariant neuron.

A potential problem with having a 3-layer representation is that the first 2 layers may contain many redundancies if both contain face space representations. Alternatively, the current first layer of the two-step model may be the intermediate stage with the capacity to contain both veridical and mirror-flipped views. The fully view-dependent representation may be prior to this and may not necessarily be comprised of any form of face space representation. But if that is the case, what is represented in the view-dependent layer? Perhaps view-specific detection, featural processing or alignment to view-specific templates. If a stage of view-dependent alignment procedures is present this may be performed in the FFA (see Baseler et al., 2016), although, this would not explain why response patterns are similar across mirror views in the FFA (Axelrod & Yovel, 2012; Flack et al., 2019; Rogers & Andrews, 2022). However, adaptation aftereffects suggest that the view-dependent representations may be coded in a face space manner. Jeffery and colleagues (2006) found aftereffects to stimuli varying in configuration and feature size, and these were larger in the observed view than the mirror view. It is possible that the results may reflect adaptation within the neurons performing alignment, but the results would perhaps be better supported by separate face space representations for each hemi-view.

Although weaker, the presence of aftereffects to mirror views even over a 90^{o} separation (Jeffery et al., 2006) also supports a layer that collates mirror views. Moreover, Jeffery et al (2007) observed a greater amount of adaptation cancellation using the mirror viewpoint than a profile viewpoint despite being separated by the same angle. If mirror views are collapsed and represented together as detailed in Chapter 5, then adaptation to a particular deviation in one view would affect the perception of its mirror due to both hemi-views being coded by a single set of neurons. The presence of a layer that collates mirror views into a single space could be tested by adapting to stimuli where the asymmetries have been accentuated. This should selectively adapt the components reflecting the asymmetries, causing the aftereffect to be asymmetrically distorted in an opponent fashion, changing the perceived hemi-view. This could be through asymmetric motion or structure. Although faces are reasonably symmetrical, there are structural and textural asymmetries such as differences in volume (Hardie et al., 2005) or the presence of moles or crooked noses as in the case of our actor.

While the mechanism behind the use of mirror views has not been fully elucidated, Chapter 5 and the current discussion provides some potential methods for representing mirror views in face space generally and within the PCA-based multi-view face space. Further work is needed to provide additional insights into how mirror views are represented and how bimodal tuning to mirror views is involved in achieving viewpoint-invariance. Future guidance can then be provided for incorporating mirror representations into the multi-view face space.

7.4 How do the models correspond to face patches in the brain?

Overall, we have provided a two-step view-invariant representation of facial motion and described how an intermediate layer collapsing mirror views might be included to form a three-step model. Here we discuss how these three stages might correspond to the face patches in the human brain. Due to the complexity of the functions and interactions of the different patches it is not possible to provide conclusive answers as to where the layers of the model would reside, but some prospective locations are discussed.

The first, fully view-dependent layer is compatible with evidence of view-dependent neurons in posterior regions of the macaque STS such as areas ML and MF (Chang & Tsao, 2017; Freiwald & Tsao, 2010; Gross & Sergent, 1992; Meyers et al., 2015; Perrett et al., 1991). It is also commensurable with evidence of view-dependence in the OFA, FFA and pSTS (Ewbank & Andrews, 2008; Fang et al., 2007; Guntupalli et al., 2017; Natu et al., 2010; Pourtois et al., 2005), although some view-invariance has been seen in these regions (Anzellotti et al., 2014; Ewbank & Andrews, 2008; Guntupalli et al., 2017; Natu et al., 2010; Ramírez et al., 2014).

The OFA may provide a good candidate for the view-dependent layer of the representation. The OFA is view-dependent (Guntupalli et al., 2017) and does not show responses to mirror views (Flack et al., 2019; Guntupalli et al., 2017; Rogers & Andrews, 2022). It is also functionally connected to the pSTS and the FFA (Davies-Thompson & Andrews, 2012; Handwerker et al., 2020), which are both sensitive to mirror views (Axelrod & Yovel, 2012; Flack et al., 2019; Rogers & Andrews, 2022). TMS of the OFA impacts both expression processing (Pitcher, 2014; Pitcher et al., 2008) and identity processing (Ambrus et al., 2017; Eick et al., 2020; Solomon-Harris et al., 2013) suggesting it can encode changeable aspects as well as information about identity. Not all evidence from TMS and fMRI suggests that the OFA is sensitive to identity (Gilaie-Dotan et al., 2010; Guntupalli et al., 2017), however, evidence from acquired prosopagnosia suggests it is crucial (Rossion et al., 2003; Schiltz et al., 2006; Steeves et al., 2006, 2007). If the OFA does process identity, then this would be compatible with identity-specific spaces. If not, then an identity-general model would still be plausible.

The OFA is thought to process features rather than global configurations (Pitcher et al., 2007) suggesting it may not represent faces globally in a face space manner. We have already noted that the fully view-dependent layer need not contain face space representations, so in that sense the OFA is suitable. However, the presence of view-specific aftereffects to different facial configurations (Jeffery et al., 2006) suggests that there is a stage where faces are processed holistically in a face space manner, which is less compatible with the feature-based processing of the OFA.

At the second level of the three-step model, representations are view-selective but invariant to mirror views. Three possible candidates for the cortical location for this layer are within the FFA, pSTS or ATL. The right FFA has been shown to process faces holistically (Rossion et al., 2000) and shows sensitivity to both viewpoint (Guntupalli et al., 2017; Weibert & Andrews, 2016) and identity (Axelrod & Yovel, 2015; Guntupalli et al., 2017; Tsantani et al., 2021; Visconti di Oleggio Castello et al., 2021). It is also the suggested homologue of macaque area ML (Tsao et al., 2003, 2008) which has been shown to use axis coding (Chang & Tsao, 2017) suggesting a face space representation in the FFA. The FFA also shows mirror-tuning (Axelrod & Yovel, 2012; Flack et al., 2019; Guntupalli et al., 2017; Rogers & Andrews, 2022).

The FFA as a potential candidate however assumes a feedforward hierarchy from the OFA, so it is unclear if it is compatible with arguments for a non-hierarchical, re-entrant model (e.g., Rossion, 2008; Solomon-Harris et al., 2013). As discussed by Rossion (2008), evidence from prosopagnosia shows patients with face-selectivity in the FFA (Rossion et al., 2003; Steeves et al., 2006) and pSTS (Sorger et al., 2007; Steeves et al., 2006) despite having no OFA, suggesting the presence of direct connections that bypass the OFA. However, the FFA in these individuals does not show a release from adaptation when a different identity is shown unlike controls (Schiltz et al., 2006; Steeves et al., 2007). Overall, they suggest that the FFA initially detects and processes the holistic configuration of the face, which may guide the OFA to finer-grained features diagnostic of identity. The OFA subsequently feeds information regarding identity back to the FFA. This re-entrant information of identity back into the FFA might support identity-specific spaces, but it is unclear if this is the case.

Although the FFA was initially proposed to process invariant aspects of faces such as identity (Haxby et al., 2000), as discussed in the introduction, there is possibly less division of labour between the FFA and pSTS than previously thought (e.g., E. Schwartz, Alreja, et al., 2023; E. Schwartz, O'Nell, et al., 2023) and numerous studies have shown that the FFA is sensitive to facial expressions (see review by Bernstein & Yovel, 2015). In their review, Bernstein and Yovel (2015) suggest that the FFA does not process the facial motion within expressions, but instead processes the facial form. We have described our model as representing dynamic information such as speech, however the components themselves represent the different forms during speech, not motion. Thus, the mirror-tuned layer could reside in the FFA. The dynamic motion of speech can be processed as the trajectories through those spaces, which could be analysed elsewhere such as in the pSTS, in conjunction with motion signals from the motion-selective area, hMT.

The pSTS may provide an alternative locus for the mirror-tuned layer. It is typically thought to be involved in processing variable and dynamic aspects of faces (Haxby et al., 2000) and is functionally connected to both the OFA and the FFA (Davies-Thompson & Andrews, 2012), so may interact with dynamic properties and identity. Decoding of identity has been reported in pSTS (Anzellotti & Caramazza, 2017; Visconti di Oleggio Castello et al., 2021), although the number of studies reporting sensitivity to identity in pSTS pales in comparison to the number of studies showing expression processing in the FFA (see review by Bernstein & Yovel, 2015). Nevertheless, the pSTS also shows mirror-tuning (Axelrod & Yovel, 2012; Flack et al., 2019; Rogers & Andrews, 2022). Consistent with learning identity-specific representations, Rogers and Andrews (2022) only found mirror-tuning for familiar faces suggesting the representation of mirror views is benefitted by prior experience with the individual. However, like the FFA, the pSTS can receive input without the OFA (Richoz et al., 2015; Sorger et al., 2007; Steeves et al., 2006) making the transition from the proposed fully view-dependent layer in OFA to the proposed mirror-tuned layer in pSTS harder to reconcile.

The ATL might also provide a locus for either the mirror-tuned layer or the multi-view layer, although the evidence towards this is even less clear. The ATL shows sensitivity to identity (Anzellotti et al., 2014; Guntupalli et al., 2017) and is the suggested homologue of macaque area AM (Rajimehr et al., 2009), which also shows axis coding of properties important for identification (2017) suggesting the ATL can hold the necessary face space representation. Although, it is worth noting that face selectivity is seen in a few clusters within the ATL, with some overlap between personally familiar and famous faces in more dorsal regions around the aSTS, particularly in the left hemisphere, and a separate ventromedial cluster that in the right hemisphere is sensitive to novel faces (Von Der Heide et al., 2013).

Interestingly, Yang and colleagues (2016) also found face selectivity in the ATL in a patient with acquired prosopagnosia missing the right FFA and right OFA. This region showed similar repetition suppression to repeated identities as controls indicating that the ATL is sensitive to identity, but it can inherit this information from regions other than the right OFA and FFA. This may be through dorsal regions such as the pSTS or the intact OFA and FFA in the left hemisphere.

The ATL is somewhat view-invariant, but is less so than the rIFG (Guntupalli et al., 2017). This intermediate view-invariance may, however, surpass mirror-tuning as Guntupalli et al found no evidence of mirror-tuning in the ATL, nor did they find clear sensitivity to viewpoint, which would still be predicted in a mirror-tuned representation. Likewise, Anzellotti et al (2014) were able to decode identity across views even when the response to mirror views was omitted, suggesting the view-invariance may surpass that of mirror-tuning. The fact that the ATL is less view-invariant than the rIFG (Guntupalli et al., 2017), however, casts doubt on whether it could instead house the multi-view, view-invariant layer of the model.

It is also unclear if the regions in ATL sensitive to identity are sensitive to facial expressions and dynamic actions. Evidence from anterior temporal lobe lobectomies have shown expression processing deficits (Milesi et al., 2014; Vliet et al., 2018) but the removed cortex included the amygdala. Other work has shown sensitivity to facial expressions in ventromedial regions near the amygdala (Hung et al., 2020), in a similar but perhaps slightly more medial region to where sensitivity to identity has been reported (Guntupalli et al., 2017; H. Yang et al., 2016). However, while emotion processing in general is widespread in the ATL (Hung et al., 2020), Avidan et al (2014) found no evidence of expression sensitivity in this slightly more lateral region. Other regions sensitive to identity have also been seen in a region slightly more anterior, dorsal and lateral (Von Der Heide et al., 2013). While some research has found a region corresponding to the ATL using dynamic localisers (H. Yang et al., 2016) not all research has (Bernstein et al., 2018). Therefore, it is unclear if the ATL provides a cortical locus for any layer of our model.

The anterior STS may be included as part of the hierarchy although currently less is known about the properties of the aSTS. The direction of head movement (left versus right) can be decoded in the right aSTS (Carlin et al., 2012) however in general the aSTS is more specialised to non-rigid deformations than the pSTS (H. Zhang et al., 2020) suggesting it may code the non-rigid deformations contained in the PCA spaces. Unlike the pSTS and mSTS, Visconti di Oleggio Castello and colleagues (2021) were unable to decode identity in the aSTS suggesting that it may process facial dynamics irrespective of identity, which would be compatible with an identity-general model of view-invariant facial motion but not identity-specific spaces. The aSTS shows evidence of coding gaze direction that is invariant to head view (Carlin et al., 2011) although only views up to ~20^{o} in yaw from frontal were used. The sensitivity to gaze could manifest through components that represent gaze, but multi-view components coding gaze would likely code direction relative to the head and not relative to the viewer, contrary to the results by Carlin et al.

The final layer of the model then forms the view-invariant layer, with each unit possessing information about multiple viewpoints. The right inferior frontal cortex poses a candidate locus of this layer. It shows view-invariance (Flack et al., 2019; Guntupalli et al., 2017) yet is sensitive to identity (Guntupalli et al., 2017), is involved in processing dynamic aspects of faces such as expressions (Nakamura et al., 1999) and is activated more by dynamic than static faces (Nikel et al., 2022). The response profile shares similarities with the pSTS, including responding equally to both visual hemifields, suggesting that the rIFG shares functional properties with the pSTS (Nikel et al., 2022) and there is evidence that they are functionally connected (Dasgupta et al., 2017; Davies-Thompson & Andrews, 2012). Thus, the IFG could collate information across the more view-dependent representations in the pSTS. The culmination of properties means it is a plausible candidate for housing the identity-specific, view-invariant representations of facial motion.

Our model is also based on representing familiar faces, yet, effects of familiarity are mixed across hemispheres. Some studies have shown effects of familiarity in right inferior frontal regions (Leveroni et al., 2000; Visconti di Oleggio Castello et al., 2021) whereas others have found effects in the left (Pourtois et al., 2005; Sun et al., 2015; Taylor et al., 2009). Rossion and colleagues (2001) found weak evidence of increased responses in the middle frontal gyrus bilaterally to familiar faces, but this was not significant. Leveroni et al (2000) found significantly larger responses in rIFG to famous faces over newly learned faces, but no difference between either compared to unfamiliar foils. Visconti di Oleggio Castello and colleagues (2021) also found higher decoding accuracy for personally familiar than newly learned identities. Of course, the effect of familiarity may be less prominent in an identity-general representation rather than the identity-specific spaces used here.

Due to the various inconsistences and caveats discussed, it is not possible to conclude exactly which face patches correspond to the different layers of the processed three-step multi-view face spaces. However, we have outlined some prospective candidates. The OFA might form the fully view-dependent stage, either the pSTS or the FFA as the intermediate stage representing mirror views, and either the ATL or more likely the IFG in representing the fully view-invariant stage.

7.5 Evidence of prototypical views

As well as elucidating which regions of cortex are involved in representing the different layers of the multi-view model, to increase the compatibility of the model with neural representations we also need to establish which views are represented within the view-dependent stage. In Chapter 6 we sought to investigate this. The models currently assume that the brain holds representations at 0°, 22.5°, 45°, 67.5° and 90°. Based on prior research in macaques (De Souza et al., 2005; Perrett et al., 1991) and behavioural results in humans (Marotta et al., 2002; O'Toole et al., 1998; Troje & Bülthoff, 1996; Van der Linde & Watson, 2010) we expected to see a decrease in thresholds needed to discriminate a veridical face from a warped face at 0°, 45° and 90°. To verify the current views used in the model, we would also expect lower thresholds for 22.5° and 67.5°.

We saw a clear inversion effect, with the threshold for inverted faces being higher than for upright faces and in general there was an increase in threshold as stimuli rotated away from frontal. Lower thresholds for generally more frontal views is consistent both with evidence that more neurons are tuned to frontal (Dubois et al., 2015; Perrett et al., 1991) and that we are exposed more to frontal views (Oruc et al., 2019).

There were not, however, lower thresholds for the expected prototypical views based on macaque and human data (0°, 45°, 90°) compared to intermediate views (22.5°, 67.5°). If anything, the thresholds were higher for these prototypical viewpoints and comparatively lower at these intermediate views. Given the neurophysiological evidence in macaques of neurons tuned to frontal, 45° and profile (De Souza et al., 2005; Perrett et al., 1991) the lower thresholds for intermediate views perhaps indicates the efficiency of view-interpolation between broad tuning curves for processing 22.5° and 67.5° views. It may also indicate the presence of previously unsampled neurons sensitive to these viewpoints. More research would need to be conducted to see if this pattern persists for different stimulus identities or different experimental paradigms.

Based on the neurophysiological evidence (De Souza et al., 2005; Perrett et al., 1991; G. Wang et al., 1996, 1998) we recommend that the multi-view model includes dedicated representations for frontal, 45° and profile. Based on the results presented here it is unclear whether dedicated representations should also be included for 22.5° and 67.5° in a model that aims to best replicate neural functioning. Ignoring neurophysiological evidence, the results presented here would suggest that it was more important to include 22.5° and 67.5° than 0°, 45° and 90°.

One way to explore whether dedicated representations for 22.5° and 67.5° views should be included is to form a model without them, and instead train the representations of the prototypical views on a wider range of angles, for example by warping 22.5° to both 0° and 45°. Likewise frontal frames could also be warped to ±45° and vice versa.

Due to the increased structural similarities across the smaller rotations, the intermediate views would likely be better reconstructed from either neighbouring prototypical space than prototypical views would be from each other's space. It is unlikely that the reconstructions of intermediate views would be as good as prototypical views in their own space, but, because intermediate views should be reconstructed reasonably well in both viewpoints, they should allow better learning of the associations between prototypical views. As such, intermediate views may also be recognised more easily because they activate two sets of neurons rather than just one.

Combined with evidence from single unit recordings in macaques, the current results generally support the use of 0°, 22.5°, 45°, 67.5° and 90° in the multi-view face space models. However, further research should be conducted to establish if the findings replicate with different stimuli and different methods to better understand why we observed higher thresholds for views defined as prototypical *a priori* and to help guide the choice of viewpoints in future developments of the multi-view face space model.

7.6 Summary

Overall, this thesis explored two aspects of linear face space. We firstly explored the response in face and object-selective cortex using fMRI when an identity-general face space was taken to the extreme. Responses were assessed for faces that varied in caricature level, ranging from average to hyper-caricatured faces beyond the realm of natural variability. The primary finding was that both the average response amplitude and response pattern consistency increased in the object-selective cortex as caricaturing increased. To our knowledge this was the first study to show that caricatured faces activate cortex typically associated with processing objects. It thus raises the interesting question of whether object-selective cortex plays any functional role in the recognition of caricatures and therefore whether hyper-caricaturing might provide a route to helping those with prosopagnosia. The response profile in the FFA and face-selective regions more generally was less clear. They showed no overall effect of caricature level in the response amplitude, however the response pattern in the right FFA became significantly more consistent as the response pattern increased. While this suggests that the FFA is sensitive to caricature level, it does not shed much light on whether an exemplar-based or norm-based face space is present in the FFA.

As well as variations in identity, within-identity variations such as changes in expression and pose can also be captured using face space representations. While much research has explored the representations of expression, little research has addressed how changes in pose might be incorporated into face space. The second line of research therefore assessed whether it was possible to form a view-invariant face space representation of facial motion based on the assumption of 2D processing. We proposed a two-step multi-view face space model which was able to reconstruct facial motion well across changes in viewpoint whilst adhering to some biological constraints, namely that the model could not be exposed to multiple, distant viewpoints at the same time. In the first layer of this model, facial expressions were coded in a view-dependent manner. The second layer combined the representations into a view-invariant layer. We then outlined a potential mechanism by which mirror views might be collapsed in such a representation, reducing computational costs while still explaining neural responses to mirror views. Overall, this model is compatible with the evidence in humans and macaques of the transition from view-dependence, to mirror-tuning and then view-invariance, and potential cortical locations for these stages were discussed. We also outlined various ways to expand the model. The model included representations at 0°, 45° and 90° based on neurophysiological results, and a further behavioural experiment supported the inclusion of intermediate representations at ~22.5° and ~67.5°. Thresholds for correctly discriminating veridical images from warped versions were comparatively lower at these intermediate viewpoints. More research will need to be conducted to see if these findings replicate, to help guide future model development.

Supplementary Materials

Supplementary figures and tables for Chapter 3 can be found in sections 3.7.2 and 3.7.3 respectively.

Supplementary videos for chapters 3, 4 and 5 can be found at the Dropbox link below. More information about the videos can be found at the end of the respective chapters (sections 3.7.1, 4.10.1 and 5.6.1) and in the “info.docx” files throughout the folders in the Dropbox.

https://www.dropbox.com/scl/fo/kaqognq18e0a5vira9syo/h?rlkey=m0k5mrhvglk69rpjdy92lk5ha&dl=0

References

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. https://doi.org/10.1016/j.cognition.2018.09.002

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3), 40–40. https://doi.org/10.1167/16.3.40

Abudarham, N., & Yovel, G. (2020). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *bioRxiv*, 2020.01.01.890277. https://doi.org/10.1101/2020.01.01.890277

Aguirre, G. K., Singh, R., & D'Esposito, M. (1999). Stimulus inversion and the responses of face and object-sensitive cortical areas. *NeuroReport*, *10*(1), 189–194.

Ahs, F., Davis, C. F., Gorka, A. X., & Hariri, A. R. (2014). Feature-based representations of emotional facial expressions in the human amygdala. *Social Cognitive and Affective Neuroscience*, *9*(9), 1372–1378. https://doi.org/10.1093/scan/nst112

Aishwarya, P., & Marcus, K. (2010). Face recognition using multiple eigenface subspaces. *Journal of Engineering and Technology Research*, *2*(8), 139–143. https://doi.org/10.5897/JETR.9000039

Ambrus, G. G., Dotzer, M., Schweinberger, S. R., & Kovács, G. (2017). The occipital face area is causally involved in the formation of identity-specific face representations. *Brain Structure and Function*, *222*(9), 4271–4282. https://doi.org/10.1007/s00429-017-1467-2

An, Z., Deng, W., Hu, J., Zhong, Y., & Zhao, Y. (2019). APA: Adaptive Pose Alignment for Pose-Invariant Face Recognition. *IEEE Access*, *7*, 14653–14670. https://doi.org/10.1109/ACCESS.2019.2894162

Anderson, N. D., & Wilson, H. R. (2005). The nature of synthetic face adaptation. *Vision Research*, *45*(14), 1815–1828. https://doi.org/10.1016/j.visres.2005.01.012

Andrews, T. J., Rogers, D., Mileva, M., Watson, D. M., Wang, A., & Burton, A. M. (2023). A narrow band of image dimensions is critical for face recognition. *Vision Research*, *212*, 108297. https://doi.org/10.1016/j.visres.2023.108297

Andrews, T. J., Watson, D. M., Rice, G. E., & Hartley, T. (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *Journal of Vision*, *15*(7), 3. https://doi.org/10.1167/15.7.3

Anzellotti, S., & Caramazza, A. (2017). Multimodal representations of person identity individuated with fMRI. *Cortex*, *89*, 85–97. https://doi.org/10.1016/j.cortex.2017.01.013

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding Representations of Face Identity That are Tolerant to Rotation. *Cerebral Cortex*, *24*(8), 1988–1995. https://doi.org/10.1093/cercor/bht046

Armann, R., Jeffery, L., Calder, A. J., & Rhodes, G. (2011). Race-specific norms for coding face identity and a functional role for norms. *Journal of Vision*, *11*(13), 9. https://doi.org/10.1167/11.13.9

Asthana, A., Marks, T. K., Jones, M. J., Tieu, K. H., & Rohith, M. (2011). Fully automatic pose-invariant face recognition via 3D pose normalization. *2011 International Conference on Computer Vision*, 937–944. https://doi.org/10.1109/ICCV.2011.6126336

Asthana, H. S., & Mandal, M. K. (1997). Hemiregional variations in facial expression of emotions. *British Journal of Psychology*, *88*(3), 519–525. https://doi.org/10.1111/j.2044-8295.1997.tb02654.x

Avidan, G., Tanzer, M., Hadj-Bouziane, F., Liu, N., Ungerleider, L. G., & Behrmann, M. (2014). Selective Dissociation Between Core and Extended Regions of the Face Processing Network in Congenital Prosopagnosia. *Cerebral Cortex*, *24*(6), 1565–1578. https://doi.org/10.1093/cercor/bht007

Axelrod, V., & Yovel, G. (2012). Hierarchical Processing of Face Viewpoint in Human Visual Cortex. *Journal of Neuroscience*, *32*(7), 2442–2452. https://doi.org/10.1523/JNEUROSCI.4770-11.2012

Axelrod, V., & Yovel, G. (2015). Successful Decoding of Famous Faces in the Fusiform Face Area. *PLOS ONE*, *10*(2), e0117126. https://doi.org/10.1371/journal.pone.0117126

Babovic-Vuksanovic, D., Messiaen, L., Nagel, C., Brems, H., Scheithauer, B., Denayer, E., Mao, R., Sciot, R., Janowski, K. M., Schuhmann, M. U., Claes, K., Beert, E., Garrity, J. A., Spinner, R. J., Stemmer-Rachamimov, A., Gavrilova, R., Van Calenbergh, F., Mautner, V., & Legius, E. (2012). Multiple orbital neurofibromas, painful peripheral nerve tumors, distinctive face and marfanoid habitus: A new syndrome. *European Journal of Human Genetics*, *20*(6), Article 6. https://doi.org/10.1038/ejhg.2011.275

Balas, B., Cox, D., & Conwell, E. (2007). The Effect of Real-World Personal Familiarity on the Speed of Face Information Processing. *PLOS ONE*, *2*(11), e1223. https://doi.org/10.1371/journal.pone.0001223

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

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), Article 7814. https://doi.org/10.1038/s41586-020-2350-5

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269

Barker, A. T., Jalinous, R., & Freeston, I. L. (1985). Non-invasive magnetic stimulation of the human motor cortex. *The Lancet*, *325*(8437), 1106–1107.

Barton, J. J. S., Cherkasova, M. V., Press, D. Z., Intriligator, J. M., & O'Connor, M. (2003). Developmental prosopagnosia: A study of three patients. *Brain and Cognition*, *51*(1), 12–30. https://doi.org/10.1016/S0278-2626(02)00516-X

Baseler, H. A., Young, A. W., Jenkins, R., Mike Burton, A., & Andrews, T. J. (2016). Face-selective regions show invariance to linear, but not to non-linear, changes in facial images. *Neuropsychologia*, *93*, 76–84. https://doi.org/10.1016/j.neuropsychologia.2016.10.004

Baudouin, J.-Y., & Gallay, M. (2006). Is face distinctiveness gender based? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 789–798. https://doi.org/10.1037/0096-1523.32.4.789

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal lobe neocortex. *Journal of Neuroscience*, *7*(2), 330–342. https://doi.org/10.1523/JNEUROSCI.07-02-00330.1987

Behrmann, M., & Williams, P. (2007). Impairments in part–whole representations of objects in two cases of integrative visual agnosia. *Cognitive Neuropsychology*, *24*(7), 701–730. https://doi.org/10.1080/02643290701672764

Benton, C. P., Jennings, S. J., & Chatting, D. J. (2006). Viewpoint dependence in adaptation to facial identity. *Vision Research*, *46*(20), 3313–3325. https://doi.org/10.1016/j.visres.2006.06.002

Berg, T., & Belhumeur, P. (2012). Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification. *In Proceeding of the British Machine Vision Conference*, *2*(7).

Beridze, J. (2021). A PCA approach to the object constancy for faces using view-based models of the face [Doctoral, UCL (University College London)]. In *Doctoral thesis, UCL (University College London).* UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10132817/

Berisha, F., Johnston, A., & McOwan, P. W. (2010). Identifying regions that carry the best information about global facial configurations. *Journal of Vision*, *10*(11), 27–27. https://doi.org/10.1167/10.11.27

Bernstein, M., Erez, Y., Blank, I., & Yovel, G. (2018). An Integrated Neural Framework for Dynamic and Static Face Processing. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-25405-9

Bernstein, M., & Yovel, G. (2015). Two neural pathways of face processing: A critical evaluation of current models. *Neuroscience & Biobehavioral Reviews*, *55*, 536–546. https://doi.org/10.1016/j.neubiorev.2015.06.010

Besle, J., Sánchez-Panchuelo, R., Bowtell, R., Francis, S., & Schluppeck, D. (2013). Event-related fMRI at 7T reveals overlapping cortical representations for adjacent fingertips in S1 of individual subjects. *Human Brain Mapping*, *35*(5), 2027–2043. https://doi.org/10.1002/hbm.22310

Beymer, D., & Poggio, T. (1995). Face recognition from one example view. *Proceedings of IEEE International Conference on Computer Vision*, 500–507. https://doi.org/10.1109/ICCV.1995.466898

Blanz, V., O'toole, Al. ice J., Vetter, T., & Wild, H. A. (2000). On The Other Side of the Mean: The Perception of Dissimilarity in Human Faces. *Perception*, *29*(8), 885–891. https://doi.org/10.1068/p2851

Blanz, V., Romdhani, S., & Vetter, T. (2002). Face identification across different poses and illuminations with a 3D morphable model. *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 202–207. https://doi.org/10.1109/AFGR.2002.1004155

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194. https://doi.org/10.1145/311535.311556

Blauch, N., & Behrmann, M. (2019). Representing faces in 3D. *Nature Human Behaviour*, *3*(8), Article 8. https://doi.org/10.1038/s41562-019-0630-6

Bona, S., Cattaneo, Z., & Silvanto, J. (2015). The Causal Role of the Occipital Face Area (OFA) and Lateral Occipital (LO) Cortex in Symmetry Perception. *Journal of Neuroscience*, *35*(2), 731–738. https://doi.org/10.1523/JNEUROSCI.3733-14.2015

Bourisly, A. K., & Shuaib, A. (2018). Neurophysiological effects of aging: A P200 ERP study. *Translational Neuroscience*, *9*(1), 61–66. https://doi.org/10.1515/tnsci-2018-0011

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

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

Bruce, V., Ness, H., Hancock, P. J. B., Newman, C., & Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, *87*(5), 894–902. https://doi.org/10.1037/0021-9010.87.5.894

Brunton, S. L., & Kutz, J. N. (2022). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press.

Bukowski, H., Dricot, L., Hanseeuw, B., & Rossion, B. (2013). Cerebral lateralization of face-sensitive areas in left-handers: Only the FFA does not get it right. *Cortex*, *49*(9), 2583–2589. https://doi.org/10.1016/j.cortex.2013.05.002

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*(1), 60–64. https://doi.org/10.1073/pnas.89.1.60

Bülthoff, I., Mohler, B. J., & Thornton, I. M. (2019). Face recognition of full-bodied avatars by active observers in a virtual environment. *Vision Research*, *157*, 242–251. https://doi.org/10.1016/j.visres.2017.12.001

Burr, D. C., Morrone, M. C., & Spinelli, D. (1989). Evidence for edge and bar detectors in human vision. *Vision Research*, *29*(4), 419–431. https://doi.org/10.1016/0042-6989(89)90006-0

Burt, A. L., & Crewther, D. P. (2020). The 4D Space-Time Dimensions of Facial Perception. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01842

Burton, A. M. (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

Burton, A. M., & Bindemann, M. (2009). The role of view in human face detection. *Vision Research*, *49*(15), 2026–2036. https://doi.org/10.1016/j.visres.2009.05.012

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., 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., & Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From

Multiple Instances. *Cognitive Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231

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

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

Butcher, N., Lander, K., Fang, H., & Costen, N. (2011). The effect of motion at encoding and retrieval for same- and other-race face recognition. *British Journal of Psychology*, *102*(4), 931–942. https://doi.org/10.1111/j.2044-8295.2011.02060.x

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision*, *17*(1), 21. https://doi.org/10.1167/17.1.21

Caharel, S., Collet, K., & Rossion, B. (2015). The early visual encoding of a face (N170) is viewpoint-dependent: A parametric ERP-adaptation study. *Biological Psychology*, *106*, 18–27. https://doi.org/10.1016/j.biopsycho.2015.01.010

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Cao, Z., Yin, Q., Tang, X., & Sun, J. (2010). Face recognition with learning-based descriptor. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2707–2714. https://doi.org/10.1109/CVPR.2010.5539992

Carbon, C.-C., & Ditye, T. (2012). Face adaptation effects show strong and long-lasting transfer from lab to more ecological contexts. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00003

Carbon, C.-C., & Leder, H. (2005). Face adaptation: Changing stable representations of familiar faces within minutes? *Advances in Experimental Psychology*, *1*(1), 1–7.

Carbon, C.-C., & Leder, H. (2006). When faces are heads: View-dependent recognition of faces altered relationally or componentially. *Swiss Journal of Psychology / Schweizerische Zeitschrift Für Psychologie / Revue Suisse de Psychologie*, *65*(4), 245–252. https://doi.org/10.1024/1421-0185.65.4.245

Carlin, J. D., Calder, A. J., Kriegeskorte, N., Nili, H., & Rowe, J. B. (2011). A Head View-Invariant Representation of Gaze Direction in Anterior Superior Temporal Sulcus. *Current Biology*, *21*(21), 1817–1821. https://doi.org/10.1016/j.cub.2011.09.025

Carlin, J. D., & Kriegeskorte, N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLOS Computational Biology*, *13*(7), e1005604. https://doi.org/10.1371/journal.pcbi.1005604

Carlin, J. D., Rowe, J. B., Kriegeskorte, N., Thompson, R., & Calder, A. J. (2012). Direction-Sensitive Codes for Observed Head Turns in Human Superior Temporal Sulcus. *Cerebral Cortex*, *22*(4), 735–744. https://doi.org/10.1093/cercor/bhr061

Cattaneo, Z., Bona, S., Monegato, M., Pece, A., Vecchi, T., Herbert, A. M., & Merabet, L. B. (2014). Visual symmetry perception in early onset monocular blindness. *Visual Cognition*, *22*(7), 963–974. https://doi.org/10.1080/13506285.2014.938712

Center, E. G., Gephart, A. M., Yang, P.-L., & Beck, D. M. (2022). Typical viewpoints of objects are better detected than atypical ones. *Journal of Vision*, *22*(12), 1. https://doi.org/10.1167/jov.22.12.1

Chai, X., Shan, S., & Gao, W. (2003). Pose normalization for robust face recognition based on statistical affine transformation. *Fourth*

International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, *3*, 1413–1417 vol.3. https://doi.org/10.1109/ICICS.2003.1292698

Chang, L., Egger, B., Vetter, T., & Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, *31*(13), 2785-2795.e4. https://doi.org/10.1016/j.cub.2021.04.014

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Chelnokova, O., & Laeng, B. (2011). Three-dimensional information in face recognition: An eye-tracking study. *Journal of Vision*, *11*(13), 27–27. https://doi.org/10.1167/11.13.27

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

Chen, D., Cao, X., Wen, F., & Sun, J. (2013). *Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification*. 3025–3032. https://www.cv-foundation.org/openaccess/content_cvpr_2013/html/Chen_Blessing_of_Dimensionality_2013_CVPR_paper.html

Chen, J., Yang, H., Wang, A., & Fang, F. (2010). Perceptual consequences of face viewpoint adaptation: Face viewpoint aftereffect, changes of differential sensitivity to face view, and their relationship. *Journal of Vision*, *10*(3), 12–12. https://doi.org/10.1167/10.3.12

Chen, W., & Liu, C. H. (2009). Transfer between pose and expression training in face recognition. *Vision Research*, *49*(3), 368–373. https://doi.org/10.1016/j.visres.2008.11.003

Choithwani, M., Almeida, S., & Egger, B. (2023). *PoseBias: On Dataset Bias and Task Difficulty - Is There an Optimal Camera Position for Facial Image Analysis?* 3096–3104. https://openaccess.thecvf.com/content/ICCV2023W/AMFG/html/Choithwani_PoseBias_On_Dataset_Bias_and_Task_Difficulty_-_Is_There_ICCVW_2023_paper.html

Chong, L. Y., Ong, T. S., & Teoh, A. B. J. (2019). Feature fusions for 2.5D face recognition in Random Maxout Extreme Learning Machine. *Applied Soft Computing*, *75*, 358–372. https://doi.org/10.1016/j.asoc.2018.11.024

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

Chung, J., & Zisserman, A. (2017). Lip reading in profile. *Ritish Machine Vision Conference, 2017*. https://ora.ox.ac.uk/objects/uuid:9f06858c-349c-416f-8ace-87751cd401fc

Colón, Y. I., Castillo, C. D., & O'Toole, A. J. (2021). Facial expression is retained in deep networks trained for face identification. *Journal of Vision*, *21*(4), 4. https://doi.org/10.1167/jov.21.4.4

Cook, M., West-Miles, J., & Chao, W. (2019). A case of prosopagnosia with the retained ability to recognize caricatures (P2.7-009). *Neurology*, *92*(15 Supplement). https://n.neurology.org/content/92/15_Supplement/P2.7-009

Cook, R., Aichelburg, C., & Johnston, A. (2015). Illusory Feature Slowing: Evidence for Perceptual Models of Global Facial Change. *Psychological Science*, *26*(4), 512–517. https://doi.org/10.1177/0956797614567340

Cook, R., Matei, M., & Johnston, A. (2011). Exploring expression space: Adaptation to orthogonal and anti-expressions. *Journal of Vision*, *11*(4), 2–2. https://doi.org/10.1167/11.4.2

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685. https://doi.org/10.1109/34.927467

Corballis, M. C., & Beale, I. L. (2020). *The Psychology of Left and Right*. Routledge.

Cowe, G. (2003). *Example-based computer-generated facial mimicry*. University College London.

Craw, I. (1992). Recognising face features and faces. *IEE Colloquium on Machine Storage and Recognition of Faces*, 7/1-7/4.

Dasgupta, S., Tyler, S. C., Wicks, J., Srinivasan, R., & Grossman, E. D. (2017). Network Connectivity of the Right STS in Three Social Perception Localizers. *Journal of Cognitive Neuroscience*, *29*(2), 221–234. https://doi.org/10.1162/jocn_a_01054

Davidenko, N., Remus, D. A., & Grill-Spector, K. (2012). Face-likeness and image variability drive responses in human face-selective ventral regions. *Human Brain Mapping*, *33*(10), 2334–2349. https://doi.org/10.1002/hbm.21367

Davies-Thompson, J., & Andrews, T. J. (2012). Intra- and interhemispheric connectivity between face-selective regions in the human brain. *Journal of Neurophysiology*, *108*(11), 3087–3095. https://doi.org/10.1152/jn.01171.2011

Davies-Thompson, J., Newling, K., & Andrews, T. J. (2013). Image-Invariant Responses in Face-Selective Regions Do Not Explain the Perceptual Advantage for Familiar Face Recognition. *Cerebral Cortex*, *23*(2), 370–377. https://doi.org/10.1093/cercor/bhs024

Davis, E. E., Matthews, C. M., & Mondloch, C. J. (2021). Ensemble coding of facial identity is not refined by experience: Evidence from other-race and inverted faces. *British Journal of Psychology*, *112*(1), 265–281. https://doi.org/10.1111/bjop.12457

Dawel, A., Wong, T. Y., McMorrow, J., Ivanovici, C., He, X., Barnes, N., Irons, J., Gradden, T., Robbins, R., Goodhew, S. C., Lane, J., & McKone, E. (2019). Caricaturing as a general method to improve poor face recognition: Evidence from low-resolution images, other-race faces, and older adults. *Journal of Experimental Psychology: Applied*, *25*(2), 256–279. https://doi.org/10.1037/xap0000180

de Gelder, B., & Rouw, R. (2000). Configural face processes in acquired and developmental prosopagnosia: Evidence for two separate face systems? *NeuroReport*, *11*(14), 3145.

de Haan, M., Johnson, M. H., Maurer, D., & Perrett, D. I. (2001). Recognition of individual faces and average face prototypes by 1- and 3-month-old infants. *Cognitive Development*, *16*(2), 659–678. https://doi.org/10.1016/S0885-2014(01)00051-X

De Souza, W. C., Eifuku, S., Tamura, R., Nishijo, H., & Ono, T. (2005). Differential Characteristics of Face Neuron Responses Within the Anterior Superior Temporal Sulcus of Macaques. *Journal of Neurophysiology*, *94*(2), 1252–1266. https://doi.org/10.1152/jn.00949.2004

Decramer, T., Premereur, E., Zhu, Q., Paesschen, W. V., Loon, J. van, Vanduffel, W., Taubert, J., Janssen, P., & Theys, T. (2021). Single-Unit Recordings Reveal the Selectivity of a Human Face Area. *Journal of Neuroscience*, *41*(45), 9340–9349. https://doi.org/10.1523/JNEUROSCI.0349-21.2021

DeGutis, J. M., Bentin, S., Robertson, L. C., & D'Esposito, M. (2007). Functional Plasticity in Ventral Temporal Cortex following Cognitive Rehabilitation of a Congenital Prosopagnosic. *Journal of Cognitive Neuroscience*, *19*(11), 1790–1802. https://doi.org/10.1162/jocn.2007.19.11.1790

Deligiannis, E., Donnelly, M., Coricelli, C., Babin, K., Stubbs, K., Ekstrand, C., Wilcox, L. M., & Culham, J. C. (2023). 3D Faces Evoke Stronger fMRI Activation than 2D Faces. *Journal of Vision*, *23*(9), 5164. https://doi.org/10.1167/jov.23.9.5164

Deng, W., Hu, J., Lu, J., & Guo, J. (2014). Transform-Invariant PCA: A Unified Approach to Fully Automatic FaceAlignment, Representation, and Recognition. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, *36*(6), 1275–1284. https://doi.org/10.1109/TPAMI.2013.194

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*(8), 2051–2062. https://doi.org/10.1523/JNEUROSCI.04-08-02051.1984

Ding, C., & Tao, D. (2016). A Comprehensive Survey on Pose-Invariant Face Recognition. *ACM Transactions on Intelligent Systems and Technology*, *7*(3), 37:1-37:42. https://doi.org/10.1145/2845089

Dobs, K., Bülthoff, I., Breidt, M., Vuong, Q. C., Curio, C., & Schultz, J. (2014). Quantifying human sensitivity to spatio-temporal information in dynamic faces. *Vision Research*, *100*, 78–87. https://doi.org/10.1016/j.visres.2014.04.009

Dobs, K., Bülthoff, I., & Schultz, J. (2016). Identity information content depends on the type of facial movement. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep34301

Dobs, K., Bülthoff, I., & Schultz, J. (2018). Use and Usefulness of Dynamic Face Stimuli for Face Perception Studies—A Review of Behavioral Findings and Methodology. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01355

Dobs, K., Ma, W. J., & Reddy, L. (2017). Near-optimal integration of facial form and motion. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-10885-y

Dobs, K., Schultz, J., Bülthoff, I., & Gardner, J. L. (2018). Task-dependent enhancement of facial expression and identity representations in human cortex. *NeuroImage*, *172*, 689–702. https://doi.org/10.1016/j.neuroimage.2018.02.013

Dolci, C., Sansone, V. A., Gibelli, D., Cappella, A., & Sforza, C. (2021). Distinctive facial features in Andersen–Tawil syndrome: A three-dimensional stereophotogrammetric analysis. *American Journal of Medical Genetics Part A*, *185*(3), 781–789. https://doi.org/10.1002/ajmg.a.62040

Dubois, J., Berker, A. O. de, & Tsao, D. Y. (2015). Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. *Journal of Neuroscience*, *35*(6), 2791–2802. https://doi.org/10.1523/JNEUROSCI.4037-14.2015

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 300–305. https://doi.org/10.1109/AFGR.1998.670965

Eger, E., Schweinberger, S. R., Dolan, R. J., & Henson, R. N. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *NeuroImage*, *26*(4), 1128–1139. https://doi.org/10.1016/j.neuroimage.2005.03.010

Eick, C. M., Kovács, G., Rostalski, S.-M., Röhrig, L., & Ambrus, G. G. (2020). The occipital face area is causally involved in identity-related visual-semantic associations. *Brain Structure and Function*, *225*(5), 1483–1493. https://doi.org/10.1007/s00429-020-02068-9

Ellis, H., & Young, A. W. (1983). The role of the right hemisphere in face perception. In *Functions of the Right Cerebral Hemisphere.* (pp. 33–64). Elsevier.

Eng, Z. H. D., Yick, Y. Y., Guo, Y., Xu, H., Reiner, M., Cham, T. J., & Chen, S. H. A. (2017). 3D faces are recognized more accurately and faster than 2D faces, but with similar inversion effects. *Vision Research*, *138*, 78–85. https://doi.org/10.1016/j.visres.2017.06.004

Epstein, R. A., Higgins, J. S., Parker, W., Aguirre, G. K., & Cooperman, S. (2006). Cortical correlates of face and scene inversion: A comparison. *Neuropsychologia*, *44*(7), 1145–1158. https://doi.org/10.1016/j.neuropsychologia.2005.10.009

Etchells, D. B., Brooks, J. L., & Johnston, R. A. (2017). Evidence for View-Invariant Face Recognition Units in Unfamiliar Face Learning. *Quarterly Journal of Experimental Psychology*, *70*(5), 874–889. https://doi.org/10.1080/17470218.2016.1248453

Ewbank, M. P., & Andrews, T. J. (2008). Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. *NeuroImage*, *40*(4), 1857–1870. https://doi.org/10.1016/j.neuroimage.2008.01.049

Faerber, S. J., Kaufmann, J. M., Leder, H., Martin, E. M., & Schweinberger, S. R. (2016). The Role of Familiarity for Representations in Norm-Based Face Space. *PLOS ONE*, *11*(5), e0155380. https://doi.org/10.1371/journal.pone.0155380

Fang, F., & He, S. (2005). Viewer-Centered Object Representation in the Human Visual System Revealed by Viewpoint Aftereffects. *Neuron*, *45*(5), 793–800. https://doi.org/10.1016/j.neuron.2005.01.037

Fang, F., Murray, S. O., & He, S. (2007). Duration-Dependent fMRI Adaptation and Distributed Viewer-Centered Face Representation in Human Visual Cortex. *Cerebral Cortex*, *17*(6), 1402–1411. https://doi.org/10.1093/cercor/bhl053

Farah, M. J., Rabinowitz, C., Quinn, G. E., & Liu, G. T. (2000). Early Commitment of Neural Substrates for Face Recognition.

Cognitive Neuropsychology, *17*(1–3), 117–123. https://doi.org/10.1080/026432900380526

Farzmahdi, A., Zarco, W., Freiwald, W. A., Kriegeskorte, N., & Golan, T. (2023). Emergence of brain-like mirror-symmetric viewpoint tuning in convolutional neural networks. *bioRxiv*. https://doi.org/10.1101/2023.01.05.522909

Favelle, S. K., Hill, H., & Claes, P. (2017). About Face: Matching Unfamiliar Faces Across Rotations of View and Lighting. *I-Perception*, *8*(6), 2041669517744221. https://doi.org/10.1177/2041669517744221

Favelle, S. K., & Palmisano, S. (2012). The Face Inversion Effect Following Pitch and Yaw Rotations: Investigating the Boundaries of Holistic Processing. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00563

Favelle, S. K., & Palmisano, S. (2018). View specific generalisation effects in face recognition: Front and yaw comparison views are better than pitch. *PLOS ONE*, *13*(12), e0209927. https://doi.org/10.1371/journal.pone.0209927

Favelle, S. K., Palmisano, S., & Avery, G. (2011). Face Viewpoint Effects about Three Axes: The Role of Configural and Featural Processing. *Perception*, *40*(7), 761–784. https://doi.org/10.1068/p6878

Favelle, S. K., Palmisano, S., & Maloney, R. T. (2007). Things are Looking up: Differential Decline in Face Recognition following Pitch and Yaw Rotation. *Perception*, *36*(9), 1334–1352. https://doi.org/10.1068/p5637

Feng, Z.-H., Huber, P., Kittler, J., Hancock, P., Wu, X.-J., Zhao, Q., Koppen, P., & Raetsch, M. (2018). Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 780–786. https://doi.org/10.1109/FG.2018.00123

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Flack, T. R., Harris, R. J., Young, A. W., & Andrews, T. J. (2019). Symmetrical Viewpoint Representations in Face-Selective Regions Convey an Advantage in the Perception and Recognition of Faces. *Journal of Neuroscience*, *39*(19), 3741–3751. https://doi.org/10.1523/JNEUROSCI.1977-18.2019

Fox, C. J., & Barton, J. J. S. (2007). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research*, *1127*, 80–89. https://doi.org/10.1016/j.brainres.2006.09.104

Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping*, *30*(5), 1637–1651. https://doi.org/10.1002/hbm.20630

Freiwald, W. A., & Hosoya, H. (2021). Neuroscience: A face's journey through space and time. *Current Biology*, *31*, R13–R15. https://doi.org/10.1016/j.cub.2020.10.065

Freiwald, W. A., & Tsao, D. Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, *330*(6005), 845–851. https://doi.org/10.1126/science.1194908

Freud, E., Ganel, T., Shelef, I., Hammer, M. D., Avidan, G., & Behrmann, M. (2017). Three-Dimensional Representations of Objects in Dorsal Cortex are Dissociable from Those in Ventral Cortex. *Cerebral Cortex*, *27*(1), 422–434. https://doi.org/10.1093/cercor/bhv229

Freunberger, R., Klimesch, W., Doppelmayr, M., & Höller, Y. (2007). Visual P2 component is related to theta phase-locking. *Neuroscience Letters*, *426*(3), 181–186. https://doi.org/10.1016/j.neulet.2007.08.062

Frowd, C. D., Hancock, P. J. B., & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites.

ACM Transactions on Applied Perception, *1*(1), 19–39. https://doi.org/10.1145/1008722.1008725

Frowd, C. D., Portch, E., Killeen, A., Mullen, L., Martin, A. J., & Hancock, P. J. B. (2019). EvoFIT Facial Composite Images: A Detailed Assessment of Impact on Forensic Practitioners, Police Investigators, Victims, Witnesses, Offenders and the Media. *2019 Eighth International Conference on Emerging Security Technologies (EST)*, 1–7. https://doi.org/10.1109/EST.2019.8806211

Furl, N., Begum, F., Sulik, J., Ferrarese, F. P., Jans, S., & Woolley, C. (2020). Face space representations of movement. *NeuroImage*, *212*, 116676. https://doi.org/10.1016/j.neuroimage.2020.116676

Furl, N., Lohse, M., & Pizzorni-Ferrarese, F. (2017). Low-frequency oscillations employ a general coding of the spatio-temporal similarity of dynamic faces. *NeuroImage*, *157*, 486–499. https://doi.org/10.1016/j.neuroimage.2017.06.023

Gad, A., Laurino, M., Maravilla, K. R., Matsushita, M., & Raskind, W. H. (2008). Sensorineural deafness, distinctive facial features, and abnormal cranial bones: A new variant of Waardenburg syndrome? *American Journal of Medical Genetics Part A*, *146A*(14), 1880–1885. https://doi.org/10.1002/ajmg.a.32402

Gao, L., Huang, Y., Zhang, Y., Zhang, X., Liu, Z., Pan, J. S., & Yu, M. (2023). Monocular information for perceiving large egocentric distance: A comparison between monocularly blind patients and normally sighted observers. *Vision Research*, *211*, 108279. https://doi.org/10.1016/j.visres.2023.108279

Gao, L., Liu, Z., Chen, Z., Pan, J. S., & Yu, M. (2023). Targeted reaching with monocular depth information and haptic feedback: Comparing between monocular patients and normally sighted observers. *Vision Research*, *211*, 108274. https://doi.org/10.1016/j.visres.2023.108274

Gardner, J., Merriam, E., Schluppeck, D., Besle, J., & Heeger, D. (2018). mrTools: Analysis and visualization package for functional magnetic resonance imaging data (Version 4.7). *Zenodo*. https://doi.org/10.5281/zenodo.1299483

Gardner, J., Sun, P., Waggoner, R. A., Ueno, K., Tanaka, K., & Cheng, K. (2005). Contrast Adaptation and Representation in Human Early Visual Cortex. *Neuron*, *47*(4), 607–620. https://doi.org/10.1016/j.neuron.2005.07.016

Garrido, L., Duchaine, B., & Nakayama, K. (2008). Face detection in normal and prosopagnosic individuals. *Journal of Neuropsychology*, *2*(1), 119–140. https://doi.org/10.1348/174866407X246843

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210. https://doi.org/10.1016/j.cognition.2010.11.002

Gilaie-Dotan, S., Gelbard-Sagiv, H., & Malach, R. (2010). Perceptual shape sensitivity to upright and inverted faces is reflected in neuronal adaptation. *NeuroImage*, *50*(2), 383–395. https://doi.org/10.1016/j.neuroimage.2009.12.077

Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. *Quarterly Journal of Experimental Psychology*, *68*(9), 1832–1843. https://doi.org/10.1080/17470218.2014.993664

Gliga, T., & Dehaene-Lambertz, G. (2007). Development of a view-invariant representation of the human head. *Cognition*, *102*(2), 261–288. https://doi.org/10.1016/j.cognition.2006.01.004

Graves, R., Goodglass, H., & Landis, T. (1982). Mouth asymmetry during spontaneous speech. *Neuropsychologia*, *20*(4), 371–381. https://doi.org/10.1016/0028-3932(82)90037-9

Griffin, H. J., McOwan, P. W., & Johnston, A. (2011). Relative faces: Encoding of family resemblance relative to gender means in face space. *Journal of Vision*, *11*(12), 8–8. https://doi.org/10.1167/11.12.8

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), Article 5. https://doi.org/10.1038/nn1224

Grill-Spector, K., Weiner, K. S., Kay, K., & Gomez, J. (2017). The Functional Neuroanatomy of Human Face Perception. *Annual Review of Vision Science*, *3*(1), 167–196. https://doi.org/10.1146/annurev-vision-102016-061214

Gross, C. G., Bender, D. B., & Mishkin, M. (1977). Contributions of the corpus callosum and the anterior commissure to visual activation of inferior temporal neurons. *Brain Research*, *131*(2), 227–239. https://doi.org/10.1016/0006-8993(77)90517-0

Gross, C. G., & Sergent, J. (1992). Face recognition. *Current Opinion in Neurobiology*, *2*(2), 156–161. https://doi.org/10.1016/0959-4388(92)90004-5

Guntupalli, J. S., Wheeler, K. G., & Gobbini, M. I. (2017). Disentangling the Representation of Identity from Head View Along the Human Face Processing Pathway. *Cerebral Cortex*, *27*(1), 46–53. https://doi.org/10.1093/cercor/bhw344

Hadjikhani, N., & de Gelder, B. (2002). Neural basis of prosopagnosia: An fMRI study. *Human Brain Mapping*, *16*(3), 176–182. https://doi.org/10.1002/hbm.10043

Halgren, E., Dale, A. M., Sereno, M. I., Tootell, R. B. H., Marinkovic, K., & Rosen, B. R. (1999). Location of human face-selective cortex with respect to retinotopic areas. *Human Brain Mapping*, *7*(1), 29–37. https://doi.org/10.1002/(SICI)1097-0193(1999)7:1<29::AID-HBM3>3.0.CO;2-R

Hancock, P. J. B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments, & Computers*, *32*(2), 327–333. https://doi.org/10.3758/BF03207802

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

Hancock, P. J. B., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network: Computation in Neural Systems*, *3*(1), 61. https://doi.org/10.1088/0954-898X/3/1/008

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

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., Somai, R. S., & Mileva, V. R. (2020). Convolutional neural net face recognition works in non-human-like ways. *Royal Society Open Science*, *7*(10), 200595. https://doi.org/10.1098/rsos.200595

Handwerker, D. A., Ianni, G., Gutierrez, B., Roopchansingh, V., Gonzalez-Castillo, J., Chen, G., Bandettini, P. A., Ungerleider, L. G., & Pitcher, D. (2020). Theta-burst TMS to the posterior superior temporal sulcus decreases resting-state fMRI connectivity across the face processing network. *Network Neuroscience*, *4*(3), 746–760. https://doi.org/10.1162/netn_a_00145

Hardie, S., Hancock, P., Rodway, P., Penton-Voak, I., Carson, D., & Wright, L. (2005). The enigma of facial asymmetry: Is there a gender-specific pattern of facedness? *Laterality*, *10*(4), 295–304. https://doi.org/10.1080/13576500442000094

Harris, A., & Aguirre, G. K. (2010). Neural Tuning for Face Wholes and Parts in Human Fusiform Gyrus Revealed by fMRI Adaptation. *Journal of Neurophysiology*, *104*(1), 336–345. https://doi.org/10.1152/jn.00626.2009

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, *75*(2), 417–429. https://doi.org/10.1007/BF00247948

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). *Effective Face Frontalization in Unconstrained Images*. 4295–4304. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Hassner_Effective_Face_Frontalization_2015_CVPR_paper.html

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The Effect of Face Inversion on Activity in Human Neural Systems for Face and Object Perception. *Neuron*, *22*(1), 189–199. https://doi.org/10.1016/S0896-6273(00)80690-X

Heilman, K. M., & Abell, T. V. D. (1980). Right hemisphere dominance for attention: The mechanism underlying hemispheric asymmetries of inattention (neglect). *Neurology*, *30*(3), 327–327. https://doi.org/10.1212/WNL.30.3.327

Hill, H., & Bruce, V. (1993). Independent Effects of Lighting, Orientation, and Stereopsis on the Hollow-Face Illusion. *Perception*, *22*(8), 887–897. https://doi.org/10.1068/p220887

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, *11*(11), 880–885. https://doi.org/10.1016/S0960-9822(01)00243-3

Hill, H., & Johnston, A. (2007). The Hollow-Face Illusion: Object-Specific Knowledge, General Assumptions or Properties of the Stimulus? *Perception*, *36*(2), 199–223. https://doi.org/10.1068/p5523

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

Hill, H., Troje, N. F., & Johnston, A. (2005). Range- and domain-specific exaggeration of facial speech. *Journal of Vision*, *5*(10), 4. https://doi.org/10.1167/5.10.4

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., Blanz, V., & O’Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), Article 11. https://doi.org/10.1038/s42256-019-0111-7

Hills, P. J., Holland, A. M., & Lewis, M. B. (2010). Aftereffects for face attributes with different natural variability: Children are more adaptable than adolescents. *Cognitive Development*, *25*(3), 278–289. https://doi.org/10.1016/j.cogdev.2010.01.002

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of Geometric Distortions on Face-Recognition Performance. *Perception*, *31*(10), 1221–1240. https://doi.org/10.1068/p3252

Hung, J., Wang, X., Wang, X., & Bi, Y. (2020). Functional subdivisions in the anterior temporal lobes: A large scale meta-analytic investigation. *Neuroscience & Biobehavioral Reviews*, *115*, 134–145. https://doi.org/10.1016/j.neubiorev.2020.05.008

Ichikawa, H., Nakato, E., Igarashi, Y., Okada, M., Kanazawa, S., Yamaguchi, M. K., & Kakigi, R. (2019). A longitudinal study of infant view-invariant face processing during the first 3–8 months of life. *NeuroImage*, *186*, 817–824. https://doi.org/10.1016/j.neuroimage.2018.11.031

Ince, R. A. A., Jaworska, K., Gross, J., Panzeri, S., van Rijsbergen, N. J., Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*, *26*(11), 4123–4135. https://doi.org/10.1093/cercor/bhw196

Ishai, A., Haxby, J. V., & Ungerleider, L. G. (2002). Visual Imagery of Famous Faces: Effects of Memory and Attention Revealed by fMRI. *NeuroImage*, *17*(4), 1729–1741. https://doi.org/10.1006/nimg.2002.1330

Isobe, S., Tamura, S., Hayamizu, S., Gotoh, Y., & Nose, M. (2021). Multi-angle lipreading using angle classification and angle-specific feature integration. *2020 International Conference on Communications, Signal Processing, and Their Applications (ICCSPA)*, 1–5. https://doi.org/10.1109/ICCSPA49915.2021.9385743

Itz, M. L., Schweinberger, S. R., & Kaufmann, J. M. (2017). Caricature generalization benefits for faces learned with enhanced idiosyncratic shape or texture. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(1), 185–197. https://doi.org/10.3758/s13415-016-0471-y

Jackson, A. S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. (2017). *Large Pose 3D Face Reconstruction From a Single Image via Direct Volumetric CNN Regression*. 1031–1039. https://openaccess.thecvf.com/content_iccv_2017/html/Jackson_Large_Pose_3D_ICCV_2017_paper.html

James, T. W., Arcurio, L. R., & Gold, J. M. (2013). Inversion Effects in Face-selective Cortex with Combinations of Face Parts. *Journal* *of Cognitive Neuroscience*, *25*(3), 455–464. https://doi.org/10.1162/jocn_a_00312

Jaquet, E., Rhodes, G., & Hayward, W. G. (2008). Race-contingent aftereffects suggest distinct perceptual norms for different race faces. *Visual Cognition*, *16*(6), 734–753. https://doi.org/10.1080/13506280701350647

Jeffery, L., Rhodes, G., & Busey, T. (2006). View-Specific Coding of Face Shape. *Psychological Science*, *17*(6), 501–505. https://doi.org/10.1111/j.1467-9280.2006.01735.x

Jeffery, L., Rhodes, G., & Busey, T. (2007). Broadly tuned, view-specific coding of face shape: Opposing figural aftereffects can be induced in different views. *Vision Research*, *47*(24), 3070–3077. https://doi.org/10.1016/j.visres.2007.08.018

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

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1888), 20181319. https://doi.org/10.1098/rspb.2018.1319

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

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jiang, F., Blanz, V., & O'Toole, A. J. (2006). Probing the Visual Representation of Faces With Adaptation: A View From the Other Side of the Mean. *Psychological Science*, *17*(6), 493–500. https://doi.org/10.1111/j.1467-9280.2006.01734.x

Jiang, F., Blanz, V., & O'Toole, A. J. (2007). The role of familiarity in three-dimensional view-transferability of face identity adaptation. *Vision Research*, *47*(4), 525–531. https://doi.org/10.1016/j.visres.2006.10.012

Jiang, F., Blanz, V., & O'Toole, A. J. (2009). Three-Dimensional Information in Face Representations Revealed by Identity Aftereffects. *Psychological Science*, *20*(3), 318–325. https://doi.org/10.1111/j.1467-9280.2009.02285.x

Johnston, A., Brown, B. B., & Elson, R. (2021). Synchronous facial action binds dynamic facial features. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-86725-x

Johnston, A., McOwan, P. W., & Benton, C. P. (1999). Robust velocity computation from a biologically motivated model of motion perception. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1418), 509–518. https://doi.org/10.1098/rspb.1999.0666

Johnston, A., McOwan, P. W., & Buxton, H. (1992). A computational model of the analysis of some first-order and second-order motion patterns by simple and complex cells. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *250*(1329), 297–306. https://doi.org/10.1098/rspb.1992.0162

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258. https://doi.org/10.1152/jn.1987.58.6.1233

Jordan, T. R., & Thomas, S. M. (2007). Hemiface contributions to hemispheric dominance in visual speech perception. *Neuropsychology*, *21*(6), 721–731. https://doi.org/10.1037/0894-4105.21.6.721

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351. https://doi.org/10.1121/1.381436

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Karimi-Rouzbahani, H., Ramezani, F., Woolgar, A., Rich, A., & Ghodrati, M. (2021). Perceptual difficulty modulates the direction of information flow in familiar face recognition. *NeuroImage*, *233*, 117896. https://doi.org/10.1016/j.neuroimage.2021.117896

Kaufmann, J. M., & Schweinberger, S. R. (2012). The faces you remember: Caricaturing shape facilitates brain processes reflecting the acquisition of new face representations. *Biological Psychology*, *89*(1), 21–33. https://doi.org/10.1016/j.biopsycho.2011.08.011

Kelly, K. R., Gallie, B. L., & Steeves, J. K. E. (2012). Impaired Face Processing in Early Monocular Deprivation from Enucleation. *Optometry and Vision Science*, *89*(2), 137–147. https://doi.org/10.1097/OPX.0b013e318240488e

Kelly, K. R., Gallie, B. L., & Steeves, J. K. E. (2019). Early monocular enucleation selectively disrupts neural development of face perception in the occipital face area. *Experimental Eye Research*, *183*, 57–61. https://doi.org/10.1016/j.exer.2018.09.013

Kietzmann, T. C., Gert, A. L., Tong, F., & König, P. (2017). Representational Dynamics of Facial Viewpoint Encoding. *Journal of Cognitive Neuroscience*, *29*(4), 637–651. https://doi.org/10.1162/jocn_a_01070

Kilgour, A. R., Kitada, R., Servos, P., James, T. W., & Lederman, S. J. (2005). Haptic face identification activates ventral occipital and temporal areas: An fMRI study. *Brain and Cognition*, *59*(3), 246–257. https://doi.org/10.1016/j.bandc.2005.07.004

Kleiner, M. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP Abstract Supplement*, 89.

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*(18), 1921–1936. https://doi.org/10.1016/S0042-6989(03)00236-0

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Knight, B., & Johnston, A. (1997). The Role of Movement in Face Recognition. *Visual Cognition*, *4*(3), 265–273. https://doi.org/10.1080/713756764

Koyano, K. W., Jones, A. P., McMahon, D. B. T., Waidmann, E. N., Russ, B. E., & Leopold, D. A. (2021). Dynamic Suppression of Average Facial Structure Shapes Neural Tuning in Three Macaque Face Patches. *Current Biology*, *31*(1), 1-12.e5. https://doi.org/10.1016/j.cub.2020.09.070

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

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*(5), 1682–1697. https://doi.org/10.1523/JNEUROSCI.3164-15.2016

Lan, Y., Theobald, B.-J., & Harvey, R. (2012). View Independent Computer Lip-Reading. *2012 IEEE International Conference on Multimedia and Expo*, 432–437. https://doi.org/10.1109/ICME.2012.192

Lander, K., & Bruce, V. (2000). Recognizing Famous Faces: Exploring the Benefits of Facial Motion. *Ecological Psychology*, *12*(4), 259–272. https://doi.org/10.1207/S15326969ECO1204_01

Lander, K., & Butcher, N. (2015). Independence of face identity and expression processing: Exploring the role of motion. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00255

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429–442. https://doi.org/10.1080/13506280444000382

Lander, K., & Davies, R. (2007). Exploring the role of characteristic motion when learning new faces. *Quarterly Journal of Experimental Psychology*, *60*(4), 519–526. https://doi.org/10.1080/17470210601117559

Lander, K., & Pitcher, D. (2017). Moving faces and moving bodies: Behavioural and neural correlates of person recognition. In *Face Processing: Systems, Disorders and Cultural Differences*.

Lane, J., Rohan, E. M. F., Sabeti, F., Essex, R. W., Maddess, T., Barnes, N., He, X., Robbins, R. A., Gradden, T., & McKone, E. (2018). Improving face identity perception in age-related macular degeneration via caricaturing. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-33543-3

Laurence, S., & Hole, G. (2012). Identity specific adaptation with composite faces. *Visual Cognition*, *20*(2), 109–120. https://doi.org/10.1080/13506285.2012.655805

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M., Kanwisher, N., Yamins, D., & DiCarlo, J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*. https://doi.org/10.1101/2020.07.09.185116

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature Effects, Distinctiveness, and Identification: Testing the Face-Space Framework. *Psychological Science*, *11*(5), 379–385. https://doi.org/10.1111/1467-9280.00274

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

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia.

Neuropsychologia, *50*(7), 1698–1707. https://doi.org/10.1016/j.neuropsychologia.2012.03.026

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), Article 7102. https://doi.org/10.1038/nature04951

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), Article 1. https://doi.org/10.1038/82947

Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research*, *69*(1), 22–29. https://doi.org/10.1007/s00426-003-0157-2

Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R., & Rao, S. M. (2000). Neural Systems Underlying the Recognition of Familiar and Newly Learned Faces. *Journal of Neuroscience*, *20*(2), 878–886. https://doi.org/10.1523/JNEUROSCI.20-02-00878.2000

Lewis, M. (2004). Face-space-R: Towards a unified account of face recognition. *Visual Cognition*, *11*(1), 29–69. https://doi.org/10.1080/13506280344000194

Li, A., Shan, S., Chen, X., & Gao, W. (2009). Maximizing intra-individual correlations for face recognition across pose differences. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 605–611. https://doi.org/10.1109/CVPR.2009.5206659

Limbach, K., Itz, M. L., Schweinberger, S. R., Jentsch, A. D., Romanova, L., & Kaufmann, J. M. (2022). Neurocognitive effects of a training program for poor face recognizers using shape and texture caricatures: A pilot investigation. *Neuropsychologia*, *165*, 108133. https://doi.org/10.1016/j.neuropsychologia.2021.108133

Little, A., Hancock, P. J. B., DeBruine, L. M., & Jones, B. C. (2012). Adaptation to Antifaces and the Perception of Correct Famous Identity in an Average Face. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00019

Liu, C. H., Bhuiyan, M. A.-A., Ward, J., & Sui, J. (2009). Transfer between pose and illumination training in face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 939. https://doi.org/10.1037/a0013710

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*(10), Article 10. https://doi.org/10.1038/nn1538

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and Physiological Evidence for Viewer-centered Object Representations in the

Primate. *Cerebral Cortex*, *5*(3), 270–288. https://doi.org/10.1093/cercor/5.3.270

Longmore, C. A., Liu, C. H., & Young, A. W. (2015). The importance of internal facial features in learning new faces. *Quarterly Journal of Experimental Psychology*, *68*(2), 249–260. https://doi.org/10.1080/17470218.2014.939666

Longmore, C. A., Lui, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77–100.

Longmore, C. A., & Tree, J. J. (2013). Motion as a cue to face recognition: Evidence from congenital prosopagnosia. *Neuropsychologia*, *51*(5), 864–875. https://doi.org/10.1016/j.neuropsychologia.2013.01.022

Lucey, P., Potamianos, G., & Sridharan, S. (2007). An Extended Pose-Invariant Lipreading System. In M. Swerts, J. Vroomem, & E. Krahmer (Eds.), *Proceedings of the Workshop on Audio-Visual Speech Processing: Cognitive and Computational Approaches* (pp. 1–5). Tilburg University. https://eprints.qut.edu.au/12843/

Marotta, J. J., McKeeff, T. J., & Behrmann, M. (2002). The effects of rotation and inversion on face processing in prosopagnosia. *Cognitive Neuropsychology*, *19*(1), 31–47. https://doi.org/10.1080/02643290143000079

Mauro, R., & Kubovy, M. (1992). Caricature and face recognition. *Memory & Cognition*, *20*(4), 433–440. https://doi.org/10.3758/BF03210927

McCleery, J. P., Zhang, L., Ge, L., Wang, Z., Christiansen, E. M., Lee, K., & Cottrell, G. W. (2008). The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence. *Vision Research*, *48*(5), 703–715. https://doi.org/10.1016/j.visres.2007.11.025

McIntire, J. P., Havig, P. R., & Geiselman, E. E. (2012). What is 3D good for? A review of human performance on stereoscopic 3D displays. *Head- and Helmet-Mounted Displays XVII; and Display Technologies and Applications for Defense, Security, and Avionics VI*, *8383*, 280–292. https://doi.org/10.1117/12.920017

McIntyre, A. H., Hancock, P. J. B., Kittler, J., & Langton, S. R. H. (2013). Improving Discrimination and Face Matching with Caricature. *Applied Cognitive Psychology*, *27*(6), 725–734. https://doi.org/10.1002/acp.2966

McKone, E., Jeffery, L., Boeing, A., Clifford, C. W. G., & Rhodes, G. (2014). Face identity aftereffects increase monotonically with adaptor extremity over, but not beyond, the range of natural faces. *Vision Research*, *98*, 1–13. https://doi.org/10.1016/j.visres.2014.01.007

Meyers, E. M., Borzello, M., Freiwald, W. A., & Tsao, D. (2015). Intelligent Information Loss: The Coding of Facial Identity, Head Pose, and Non-Face Information in the Macaque Face Patch System. *Journal of Neuroscience*, *35*(18), 7069–7081. https://doi.org/10.1523/JNEUROSCI.3086-14.2015

Milborrow, S., Bishop, T., & Nicolls, F. (2013). *Multiview Active Shape Models with SIFT Descriptors for the 300-W Face Landmark Challenge*. 378–385. https://www.cv-foundation.org/openaccess/content_iccv_workshops_2013/W11/html/Milborrow_Multiview_Active_Shape_2013_ICCV_paper.html

Milesi, V., Cekic, S., Péron, J., Frühholz, S., Cristinzio, C., Seeck, M., & Grandjean, D. (2014). Multimodal emotion perception after anterior temporal lobectomy (ATL). *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnhum.2014.00275

Mileva, M., & Burton, A. M. (2019). Face search in CCTV surveillance. *Cognitive Research: Principles and Implications*, *4*(1), 37. https://doi.org/10.1186/s41235-019-0193-0

Minnebusch, D. A., Suchan, B., Köster, O., & Daum, I. (2009). A bilateral occipitotemporal network mediates face perception. *Behavioural Brain Research*, *198*(1), 179–185. https://doi.org/10.1016/j.bbr.2008.10.041

Minnebusch, D. A., Suchan, B., Ramon, M., & Daum, I. (2007). Event-related potentials reflect heterogeneity of developmental prosopagnosia. *European Journal of Neuroscience*, *25*(7), 2234–2247. https://doi.org/10.1111/j.1460-9568.2007.05451.x

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with Links: A Unified System for Processing Faces in the Macaque Temporal Lobe. *Science*, *320*(5881), 1355–1359. https://doi.org/10.1126/science.1157436

Mondloch, C. J., Davis, E., & Matthews, C. (2023, April 19). *Ensemble coding of facial identity is robust, but an unlikely route to face learning* [Conference Presentation]. Experimental Psychology Meeting, University of Plymouth, Plymouth, UK.

Morris, J. S., Friston, K. J., Büchel, C., Frith, C. D., Young, A. W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain*, *121*(1), 47–57. https://doi.org/10.1093/brain/121.1.47

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, *383*(6603), Article 6603. https://doi.org/10.1038/383812a0

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

Murphy, J., Ipser, A., Gaigg, S., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577–581. https://doi.org/10.1037/xhp0000049

Murty, R. N. A., & Arun, S. P. (2015). Dynamics of 3D view invariance in monkey inferotemporal cortex. *Journal of Neurophysiology*, *113*(7), 2180–2194. https://doi.org/10.1152/jn.00810.2014

Murty, R. N. A., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, *117*(37), 23011–23020. https://doi.org/10.1073/pnas.2004607117

Nagle, F., Griffin, H., Johnston, A., & McOwan, P. (2013). Techniques for Mimicry and Identity Blending Using Morph Space PCA. In J.-I. Park & J. Kim (Eds.), *Computer Vision—ACCV 2012 Workshops* (pp. 296–307). Springer. https://doi.org/10.1007/978-3-642-37484-5_25

Nakamura, K., Kawashima, R., Ito, K., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., & Kojima, S.

(1999). Activation of the Right Inferior Frontal Cortex During Assessment of Facial Emotion. *Journal of Neurophysiology*, *82*(3), 1610–1614. https://doi.org/10.1152/jn.1999.82.3.1610

Nasr, S., & Tootell, R. B. H. (2012). Role of fusiform and anterior temporal cortical areas in facial recognition. *NeuroImage*, *63*(3), 1743–1753. https://doi.org/10.1016/j.neuroimage.2012.08.031

Natale, M., Gur, R. E., & Gur, R. C. (1983). Hemispheric asymmetries in processing emotional expressions. *Neuropsychologia*, *21*(5), 555–565. https://doi.org/10.1016/0028-3932(83)90011-8

Natu, V. S., Jiang, F., Narvekar, A., Keshvari, S., Blanz, V., & O'Toole, A. J. (2010). Dissociable Neural Patterns of Facial Identity across Changes in Viewpoint. *Journal of Cognitive Neuroscience*, *22*(7), 1570–1582. https://doi.org/10.1162/jocn.2009.21312

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

Nicholls, M. E. R., Ellis, B. E., Clement, J. G., & Yoshino, M. (2004). Detecting hemifacial asymmetries in emotional expression with three–dimensional computerized image analysis. *Proceedings of* *the Royal Society of London. Series B: Biological Sciences*, *271*(1540), 663–668. https://doi.org/10.1098/rspb.2003.2660

Nicholls, M. E. R., & Searle, D. A. (2006). Asymmetries for the visual expression and perception of speech. *Brain and Language*, *97*(3), 322–331. https://doi.org/10.1016/j.bandl.2005.11.007

Nikel, L., Sliwinska, M. W., Kucuk, E., Ungerleider, L. G., & Pitcher, D. (2022). *Measuring the response to visually presented faces in the human lateral prefrontal cortex* (p. 2022.03.06.483119). bioRxiv. https://doi.org/10.1101/2022.03.06.483119

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

O’Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), Article 6753. https://doi.org/10.1038/44134

O’Craven, K. M., & Kanwisher, N. (2000). Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions. *Journal of Cognitive Neuroscience*, *12*(6), 1013–1023. https://doi.org/10.1162/08989290051137549

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly*, *25*(5), 817–834. https://doi.org/10.1016/j.leaqua.2014.06.002

Oruc, I., Shafai, F., Murthy, S., Lages, P., & Ton, T. (2019). The adult face-diet: A naturalistic observation study. *Vision Research*, *157*, 222–229. https://doi.org/10.1016/j.visres.2018.01.001

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, *38*(15), 2351–2363. https://doi.org/10.1016/S0042-6989(98)00042-X

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, *6*(6), 261–266. https://doi.org/10.1016/S1364-6613(02)01908-3

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition*, *170*, 190–200. https://doi.org/10.1016/j.cognition.2017.10.002

Palermo, R., Rivolta, D., Wilson, C. E., & Jeffery, L. (2011). Adaptive face space coding in congenital prosopagnosia: Typical figural aftereffects but abnormal identity aftereffects. *Neuropsychologia*, *49*(14), 3801–3812. https://doi.org/10.1016/j.neuropsychologia.2011.09.039

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*. https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1

Parr, L. A., Taubert, J., Little, A. C., & Hancock, P. J. B. (2012). The organization of conspecific face space in nonhuman primates. *The Quarterly Journal of Experimental Psychology*, *65*(12), 2411–2434. https://doi.org/10.1080/17470218.2012.693110

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S., & Grill-Spector, K. (2012). Electrical Stimulation of Human Fusiform Face-Selective Regions Distorts Face Perception. *Journal of Neuroscience*, *32*(43), 14915–14920. https://doi.org/10.1523/JNEUROSCI.2609-12.2012

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pelli, D. G. (1997). *The VideoToolbox software for visual psychophysics: Transforming numbers into movies*. *10*(4), 437–442.

Pentland, Moghaddam, & Starner. (1994). View-based and modular eigenspaces for face recognition. *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 84–91. https://doi.org/10.1109/CVPR.1994.323814

Perrett, D. I., Mistlin, A. J., Chitty, A. J., Smith, P. A. J., Potter, D. D., Broennimann, R., & Harries, M. (1988). Specialized face processing and hemispheric asymmetry in man and monkey: Evidence from single unit and reaction time studies. *Behavioural Brain Research*, *29*(3), 245–258. https://doi.org/10.1016/0166-4328(88)90029-0

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Experimental Brain Research*, *86*(1), 159–173. https://doi.org/10.1007/BF00231050

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II: Effects of attention. *Electroencephalography and Clinical Neurophysiology*, *36*, 191–200. https://doi.org/10.1016/0013-4694(74)90156-4

Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing Moving Faces: The Relative Contribution of Motion and Perspective View Information. *Visual Cognition*, *4*(4), 409–438. https://doi.org/10.1080/713756769

Pilz, K. S., Thornton, I. M., & Bülthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research*, *171*(4), 436–447. https://doi.org/10.1007/s00221-005-0283-8

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009). Neural Representations of Faces and Body Parts in Macaque and Human Cortex: A Comparative fMRI Study. *Journal of Neurophysiology*, *101*(5), 2581–2600. https://doi.org/10.1152/jn.91198.2008

Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences*, *102*(19), 6996–7001. https://doi.org/10.1073/pnas.0502605102

Pitcher, D. (2014). Facial Expression Recognition Takes Longer in the Posterior Superior Temporal Sulcus than in the Occipital Face Area. *Journal of Neuroscience*, *34*(27), 9173–9177. https://doi.org/10.1523/JNEUROSCI.5038-13.2014

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, *56*(4), 2356–2363. https://doi.org/10.1016/j.neuroimage.2011.03.067

Pitcher, D., Duchaine, B., & Walsh, V. (2014). Combined TMS and fMRI Reveal Dissociable Cortical Pathways for Dynamic and Static Face Perception. *Current Biology*, *24*(17), 2066–2070. https://doi.org/10.1016/j.cub.2014.07.060

Pitcher, D., Garrido, L., Walsh, V., & Duchaine, B. C. (2008). Transcranial Magnetic Stimulation Disrupts the Perception and Embodiment of Facial Expressions. *Journal of Neuroscience*, *28*(36), 8929–8933. https://doi.org/10.1523/JNEUROSCI.1450-08.2008

Pitcher, D., Ianni, G., Holiday, K., & Ungerleider, L. G. (2023). Identifying the cortical face network with dynamic face stimuli: A large group fMRI study. *bioRxiv*. https://doi.org/10.1101/2023.09.26.559583

Pitcher, D., Japee, S., Rauth, L., & Ungerleider, L. G. (2017). The Superior Temporal Sulcus Is Causally Connected to the Amygdala: A Combined TBS-fMRI Study. *Journal of Neuroscience*, *37*(5), 1156–1161. https://doi.org/10.1523/JNEUROSCI.0114-16.2016

Pitcher, D., Pilkington, A., Rauth, L., Baker, C., Kravitz, D. J., & Ungerleider, L. G. (2020). The Human Posterior Superior Temporal Sulcus Samples Visual Space Differently From Other Face-Selective Regions. *Cerebral Cortex*, *30*(2), 778–785. https://doi.org/10.1093/cercor/bhz125

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS Evidence for the Involvement of the Right Occipital Face Area in Early Face Processing. *Current Biology*, *17*(18), 1568–1573. https://doi.org/10.1016/j.cub.2007.07.063

Polosecki, P., Moeller, S., Schweers, N., Romanski, L. M., Tsao, D. Y., & Freiwald, W. A. (2013). Faces in Motion: Selectivity of Macaque and Human Face Processing Areas for Dynamic Stimuli. *Journal of Neuroscience*, *33*(29), 11768–11773. https://doi.org/10.1523/JNEUROSCI.5402-11.2013

Pourtois, G., Schwartz, S., Seghier, M. L., Lazeyras, F., & Vuilleumier, P. (2005). View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: An event-related fMRI study. *NeuroImage*, *24*(4), 1214–1224. https://doi.org/10.1016/j.neuroimage.2004.10.038

Powell, J., Letson, S., Davidoff, J., Valentine, T., & Greenwood, R. (2008). Enhancement of face recognition learning in patients with brain injury using three cognitive training procedures. *Neuropsychological Rehabilitation*, *18*(2), 182–203. https://doi.org/10.1080/09602010701419485

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential Sensitivity of Human Visual Cortex to Faces, Letterstrings, and Textures: A Functional Magnetic Resonance Imaging Study. *Journal of Neuroscience*, *16*(16), 5205–5215. https://doi.org/10.1523/JNEUROSCI.16-16-05205.1996

Rajimehr, R., Young, J. C., & Tootell, R. B. H. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences*, *106*(6), 1995–2000. https://doi.org/10.1073/pnas.0807304106

Ramírez, F. M. (2018). Orientation Encoding and Viewpoint Invariance in Face Recognition: Inferring Neural Properties from Large-Scale Signals. *The Neuroscientist*, *24*(6), 582–608. https://doi.org/10.1177/1073858418769554

Ramírez, F. M., Cichy, R. M., Allefeld, C., & Haynes, J.-D. (2014). The Neural Code for Face Orientation in the Human Fusiform Face Area. *Journal of Neuroscience*, *34*(36), 12155–12167. https://doi.org/10.1523/JNEUROSCI.3156-13.2014

Ramon, M., Caharel, S., & Rossion, B. (2011). The Speed of Recognition of Personally Familiar Faces. *Perception*, *40*(4), 437–449. https://doi.org/10.1068/p6794

Ramon, M., Dricot, L., & Rossion, B. (2010). Personally familiar faces are perceived categorically in face-selective regions other than the fusiform face area. *European Journal of Neuroscience*, *32*(9), 1587–1598. https://doi.org/10.1111/j.1460-9568.2010.07405.x

Ramon, M., & Rossion, B. (2010). Impaired processing of relative distances between features and of the eye region in acquired prosopagnosia—Two sides of the same holistic coin? *Cortex*, *46*(3), 374–389. https://doi.org/10.1016/j.cortex.2009.06.001

Rangarajan, V., Hermes, D., Foster, B. L., Weiner, K. S., Jacques, C., Grill-Spector, K., & Parvizi, J. (2014). Electrical Stimulation of the Left and Right Human Fusiform Gyrus Causes Different Effects in Conscious Face Perception. *Journal of Neuroscience*, *34*(38), 12828–12836. https://doi.org/10.1523/JNEUROSCI.0527-14.2014

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, *19*(4), 473–497. https://doi.org/10.1016/0010-0285(87)90016-8

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977–2987. https://doi.org/10.1016/j.visres.2006.03.002

Rhodes, G., Pond, S., Burton, N., Kloth, N., Jeffery, L., Bell, J., Ewing, L., Calder, A. J., & Palermo, R. (2015). How distinct is the coding of face identity and expression? Evidence for some common dimensions in face space. *Cognition*, *142*, 123–137. https://doi.org/10.1016/j.cognition.2015.05.012

Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., & McKay, R. (2001). Do facial averageness and symmetry signal health? *Evolution and Human Behavior*, *22*(1), 31–46. https://doi.org/10.1016/S1090-5138(00)00060-X

Richoz, A.-R., Jack, R. E., Garrod, O. G. B., Schyns, P. G., & Caldara, R. (2015). Reconstructing dynamic mental models of facial expressions in prosopagnosia reveals distinct representations for identity and expression. *Cortex*, *65*, 50–64. https://doi.org/10.1016/j.cortex.2014.11.015

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., Mireku, M. O., & Kramer, R. S. S. (2020). Face averages and multiple images in a live matching task. *British Journal of Psychology*, *111*(1), 92–102. https://doi.org/10.1111/bjop.12388

Rivest, J., Moscovitch, M., & Black, S. (2009). A comparative case study of face recognition: The contribution of configural and part-based recognition systems, and their interaction. *Neuropsychologia*, *47*(13), 2798–2811. https://doi.org/10.1016/j.neuropsychologia.2009.06.004

Robbins, R., McKone, E., & Edwards, M. (2007). Aftereffects for face attributes with different natural variability: Adapter position effects and neural models. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 570–592. https://doi.org/10.1037/0096-1523.33.3.570

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

Rodríguez, J., Bortfeld, H., Rudomín, I., Hernández, B., & Gutiérrez-Osuna, R. (2009). The reverse-caricature effect revisited: Familiarization with frontal facial caricatures improves veridical face recognition. *Applied Cognitive Psychology*, *23*(5), 733–742. https://doi.org/10.1002/acp.1539

Rogers, D., & Andrews, T. J. (2022). The emergence of view-symmetric neural responses to familiar and unfamiliar faces. *Neuropsychologia*, *172*, 108275. https://doi.org/10.1016/j.neuropsychologia.2022.108275

Rolls, E. T., Baylis, G. C., Hasselmo, M. E., & Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, *76*(1), 153–164. https://doi.org/10.1007/BF00253632

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics*, *68*(1), 84–93. https://doi.org/10.3758/BF03193658

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, *64*(2), 220–229. https://doi.org/10.3758/BF03195788

Ross, D. A., Hancock, P. J. B., & Lewis, M. B. (2010). Changing faces: Direction is important. *Visual Cognition*, *18*(1), 67–81. https://doi.org/10.1080/13506280802536656

Rossion, B. (2008). Constraining the cortical face network by neuroimaging studies of acquired prosopagnosia. *NeuroImage*, *40*(2), 423–426. https://doi.org/10.1016/j.neuroimage.2007.10.047

Rossion, B., Caldara, R., Seghier, M., Schuller, A., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, *126*(11), 2381–2395. https://doi.org/10.1093/brain/awg241

Rossion, B., Dricot, L., Devolder, A., Bodart, J.-M., Crommelinck, M., Gelder, B. de, & Zoontjes, R. (2000). Hemispheric Asymmetries for Whole-Based and Part-Based Face Processing in the Human Fusiform Gyrus. *Journal of Cognitive Neuroscience*, *12*(5), 793–802. https://doi.org/10.1162/089892900562606

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How Does the Brain Discriminate Familiar and Unfamiliar Faces?: A PET Study of Face Categorical Perception. *Journal of*

Cognitive Neuroscience, *13*(7), 1019–1034. https://doi.org/10.1162/089892901753165917

Rossion, B., & Taubert, J. (2019). What can we learn about human individual face recognition from experimental studies in monkeys? *Vision Research*, *157*, 142–158. https://doi.org/10.1016/j.visres.2018.03.012

Royer, J., Blais, C., Barnabé-Lortie, V., Carré, M., Leclerc, J., & Fiset, D. (2016). Efficient visual information for unfamiliar face matching despite viewpoint variations: It's not in the eyes! *Vision Research*, *123*, 33–40. https://doi.org/10.1016/j.visres.2016.04.004

Royer, J., Blais, C., Charbonneau, I., Déry, K., Tardif, J., Duchaine, B., Gosselin, F., & Fiset, D. (2018). Greater reliance on the eye region predicts better face recognition ability. *Cognition*, *181*, 12–20. https://doi.org/10.1016/j.cognition.2018.08.004

Rupnik, J., & Shawe-Taylor, J. (2010). *Multi-view canonical correlation analysis*. In conference on data mining and data warehouses (SiKDD 2010).

Russell, R., & Sinha, P. (2007). Real-World Face Recognition: The Importance of Surface Reflectance Properties. *Perception*, *36*(9), 1368–1374. https://doi.org/10.1068/p5779

Ryu, J.-J., & Chaudhuri, A. (2006). Representations of familiar and unfamiliar faces as revealed by viewpoint-aftereffects. *Vision Research*, *46*(23), 4059–4063. https://doi.org/10.1016/j.visres.2006.07.018

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, *48*(12), 3596–3605. https://doi.org/10.1016/j.neuropsychologia.2010.08.009

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

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

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

Schiltz, C., Sorger, B., Caldara, R., Ahmed, F., Mayer, E., Goebel, R., & Rossion, B. (2006). Impaired Face Discrimination in Acquired Prosopagnosia Is Associated with Abnormal Response to

Individual Faces in the Right Middle Fusiform Gyrus. *Cerebral Cortex*, *16*(4), 574–586. https://doi.org/10.1093/cercor/bhj005

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is It Really Robust? *Methodology*, *6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Scholes, C., Skipper, J. I., & Johnston, A. (2020). The interrelationship between the face and vocal tract configuration during audiovisual speech. *Proceedings of the National Academy of Sciences*, *117*(51), 32791–32798. https://doi.org/10.1073/pnas.2006192117

Schrouff, J., Raccah, O., Baek, S., Rangarajan, V., Salehi, S., Mourão-Miranda, J., Helili, Z., Daitch, A. L., & Parvizi, J. (2020). Fast temporal dynamics and causal relevance of face processing in the human temporal cortex. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14432-8

Schulz, C., Kaufmann, J. M., Walther, L., & Schweinberger, S. R. (2012). Effects of anticaricaturing vs. Caricaturing and their neural correlates elucidate a role of shape for face learning. *Neuropsychologia*, *50*(10), 2426–2434. https://doi.org/10.1016/j.neuropsychologia.2012.06.013

Schwaninger, A., Schumacher, S., Bülthoff, H., & Wallraven, C. (2007). Using 3D computer graphics for perception: The role of local and global information in face processing. *Proceedings of the 4th*

Symposium on Applied Perception in Graphics and Visualization, 19–26. https://doi.org/10.1145/1272582.1272586

Schwartz, E., Alreja, A., Richardson, R. M., Ghuman, A., & Anzellotti, S. (2023). Intracranial Electroencephalography and Deep Neural Networks Reveal Shared Substrates for Representations of Face Identity and Expressions. *Journal of Neuroscience*, *43*(23), 4291–4303. https://doi.org/10.1523/JNEUROSCI.1277-22.2023

Schwartz, E., O'Nell, K., Saxe, R., & Anzellotti, S. (2023). Challenging the Classical View: Recognition of Identity and Expression as Integrated Processes. *Brain Sciences*, *13*(2), Article 2. https://doi.org/10.3390/brainsci13020296

Schwartz, L., & Yovel, G. (2016). The roles of perceptual and conceptual information in face recognition. *Journal of Experimental Psychology: General*, *145*(11), 1493–1511. https://doi.org/10.1037/xge0000220

Schweinberger, S. R., & Burton, A. M. (2003). Covert Recognition and the Neural System for Face Processing. *Cortex*, *39*(1), 9–30. https://doi.org/10.1016/S0010-9452(08)70071-6

Schweinberger, S. R., Pickering, E. C., Jentzsch, I., Burton, A. M., & Kaufmann, J. M. (2002). Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Cognitive Brain Research*, *14*(3), 398–409. https://doi.org/10.1016/S0926-6410(02)00142-8

Sergent, J., & Signoret, J.-L. (1992). Varieties of Functional Deficits in Prosopagnosia. *Cerebral Cortex*, *2*(5), 375–388. https://doi.org/10.1093/cercor/2.5.375

Shan, S., Gao, W., & Zhao, D. (2003). Face recognition based on face-specific subspace. *International Journal of Imaging Systems and Technology*, *13*(1), 23–32. https://doi.org/10.1002/ima.10047

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics*, *66*(2), 352–362. https://doi.org/10.3758/BF03194884

Shirakata, T., & Saitoh, T. (2020). *Lip Reading using Facial Expression Features*.

Short, L. A., Hatry, A. J., & Mondloch, C. J. (2011). The development of norm-based coding and race-specific face prototypes: An examination of 5- and 8-year-olds’ face space. *Journal of Experimental Child Psychology*, *108*(2), 338–357. https://doi.org/10.1016/j.jecp.2010.07.007

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *JOSA A*, *4*(3), 519–524. https://doi.org/10.1364/JOSAA.4.000519

Skinner, A. L., & Benton, C. P. (2012). The expressions of strangers: Our identity-independent representation of facial expression. *Journal of Vision*, *12*(2), 12. https://doi.org/10.1167/12.2.12

Sliwinska, M. W., Bearpark, C., Corkhill, J., McPhillips, A., & Pitcher, D. (2020). Dissociable pathways for moving and static face perception begin in early visual cortex: Evidence from an acquired prosopagnosic. *Cortex*, *130*, 327–339. https://doi.org/10.1016/j.cortex.2020.03.033

Sliwinska, M. W., Brown, L., Earl, M., O’Gorman, D., Pollicina, G., Burton, M., & Pitcher, D. (2019). *Face learning via short real-world social interactions induces changes in face-selective brain areas and hippocampus*. PsyArXiv. https://doi.org/10.31234/osf.io/hw27y

Solomon-Harris, L. M., Mullin, C. R., & Steeves, J. K. E. (2013). TMS to the “occipital face area” affects recognition but not categorization of faces. *Brain and Cognition*, *83*(3), 245–251. https://doi.org/10.1016/j.bandc.2013.08.007

Somai, R. S., & Hancock, P. J. B. (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

Sorger, B., Goebel, R., Schiltz, C., & Rossion, B. (2007). Understanding the functional neuroanatomy of acquired prosopagnosia. *NeuroImage*, *35*(2), 836–852. https://doi.org/10.1016/j.neuroimage.2006.09.051

Steeves, J., Culham, J. C., Duchaine, B. C., Pratesi, C. C., Valyear, K. F., Schindler, I., Humphrey, G. K., Milner, A. D., & Goodale, M. A. (2006). The fusiform face area is not sufficient for face recognition: Evidence from a patient with dense prosopagnosia and no occipital face area. *Neuropsychologia*, *44*(4), 594–609. https://doi.org/10.1016/j.neuropsychologia.2005.06.013

Steeves, J., Goltz, H., Dricot, L., Sorger, B., Peters, J., Milner, A. D., Goodale, M., & Rossion, B. (2007). Face-selective activation in the middle fusiform gyrus in a patient with acquired prosopagnosia: Abnormal modulation for face identity. *Journal of Vision*, *7*(9), 627. https://doi.org/10.1167/7.9.627

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12), 6. https://doi.org/10.1167/14.12.6

Sun, D., Lee, T. M. C., & Chan, C. C. H. (2015). Unfolding the Spatial and Temporal Neural Processing of Lying about Face Familiarity. *Cerebral Cortex*, *25*(4), 927–936. https://doi.org/10.1093/cercor/bht284

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

Susilo, T., McKone, E., & Edwards, M. (2010). What shape are the neural response functions underlying opponent coding in face space? A psychophysical investigation. *Vision Research*, *50*(3), 300–314. https://doi.org/10.1016/j.visres.2009.11.016

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, N. S. (1961). Figural after-effects and apparent size. *Quarterly Journal of Experimental Psychology*, *13*(4), 222–228. https://doi.org/10.1080/17470216108416498

Svart, N., & Starrfelt, R. (2022). Is It Just Face Blindness? Exploring Developmental Comorbidity in Individuals with Self-Reported Developmental Prosopagnosia. *Brain Sciences*, *12*(2), Article 2. https://doi.org/10.3390/brainsci12020230

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

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). *DeepFace: Closing the Gap to Human-Level Performance in Face Verification*. 1701–1708. https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Taigman_DeepFace_Closing_the_2014_CVPR_paper.html

Tanaka, J. W., & Pierce, L. J. (2009). The neural plasticity of other-race face recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 122–131. https://doi.org/10.3758/CABN.9.1.122

Taubert, J., Japee, S., Murphy, A. P., Tardiff, C. T., Koele, E. A., Kumar, S., Leopold, D. A., & Ungerleider, L. G. (2020). Parallel Processing of Facial Expression and Head Orientation in the Macaque Brain. *Journal of Neuroscience*, *40*(42), 8119–8131. https://doi.org/10.1523/JNEUROSCI.0524-20.2020

Taylor, M. J., Arsalidou, M., Bayless, S. J., Morris, D., Evans, J. W., & Barbeau, E. J. (2009). Neural correlates of personally familiar faces: Parents, partner and own faces. *Human Brain Mapping*, *30*(7), 2008–2020. https://doi.org/10.1002/hbm.20646

Theobald, B. J., Harvey, R., Cox, S. J., Lewis, C., & Owen, G. P. (2006). Lip-reading enhancement for law enforcement. *Optics and Photonics for Counterterrorism and Crime Fighting II*, *6402*, 24–32. https://doi.org/10.1117/12.689960

Thornton, I. M., & Kourtzi, Z. (2002). A Matching Advantage for Dynamic Human Faces. *Perception*, *31*(1), 113–132. https://doi.org/10.1068/p3300

Thornton, I. M., Mullins, E., & Banahan, K. (2011). Motion can amplify the face-inversion effect. *Psihologija*, *44*(1), 5–22.

Tran, L., Yin, X., & Liu, X. (2017). *Disentangled Representation Learning GAN for Pose-Invariant Face Recognition*. 1415–1424. https://openaccess.thecvf.com/content_cvpr_2017/html/Tran_Disentangled_Representation_Learning_CVPR_2017_paper.html

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, *36*(12), 1761–1771. https://doi.org/10.1016/0042-6989(95)00230-8

Troje, N. F., & Kersten, D. (1999). Viewpoint-Dependent Recognition of Familiar Faces. *Perception*, *28*(4), 483–487. https://doi.org/10.1068/p2901

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA Encode Distinct Types of

Face Identity Information. *Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020

Tsao, D. Y. (2006). A Dedicated System for Processing Faces. *Science*, *314*(5796), 72–73. https://doi.org/10.1126/science.1135163

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, *6*(9), Article 9. https://doi.org/10.1038/nn1111

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, *105*(49), 19514–19519. https://doi.org/10.1073/pnas.0809662105

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Valentin, D., Abdi, H., O'Toole, A. J., & Cottrell, G. W. (1994). Connectionist models of face processing: A survey. *Pattern Recognition*, *27*(9), 1209–1230. https://doi.org/10.1016/0031-3203(94)90006-X

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., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, *69*(10), 1996–2019. https://doi.org/10.1080/17470218.2014.990392

Van der Linde, I., & Watson, T. (2010). A combinatorial study of pose effects in unfamiliar face recognition. *Vision Research*, *50*(5), 522–533. https://doi.org/10.1016/j.visres.2009.12.012

Vetter, T., Poggio, T., & Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, *4*(1), 18–23. https://doi.org/10.1016/S0960-9822(00)00004-X

Vignal, J. P., Chauvel, P., & Halgren, E. (2000). Localised Face Processing by the Human Prefrontal Cortex: Stimulation-Evoked Hallucinations of Faces. *Cognitive Neuropsychology*, *17*(1–3), 281–291. https://doi.org/10.1080/026432900380616

Visconti di Oleggio Castello, M., Haxby, J. V., & Gobbini, M. I. (2021). Shared neural codes for visual and semantic information about familiar faces in a common representational space. *Proceedings of the National Academy of Sciences*, *118*(45), e2110474118. https://doi.org/10.1073/pnas.2110474118

Vliet, L. V. de, Jastorff, J., Huang, Y.-A., Paesschen, W. V., Vandenbulcke, M., & Stock, J. V. den. (2018). Anterior Temporal Lobectomy Impairs Neural Classification of Body Emotions in Right Superior Temporal Sulcus and Reduces Emotional Enhancement in Distributed Brain Areas without Affecting Behavioral Classification. *Journal of Neuroscience*, *38*(43), 9263–9274. https://doi.org/10.1523/JNEUROSCI.0634-18.2018

Von Der Heide, R., Skipper, L., & Olson, I. (2013). Anterior temporal face patches: A meta-analysis and empirical study. *Frontiers in Human Neuroscience*, *7*. https://www.frontiersin.org/articles/10.3389/fnhum.2013.00017

Vuong, Q. C., & Tarr, M. (n.d.). *Do we remember the motions of objects*. Retrieved October 18, 2023, from https://www.researchgate.net/profile/Quoc-Vuong/publication/2558090_Do_We_Remember_the_Motions_of_Objects/links/0046351af2081ac634000000/Do-We-Remember-the-Motions-of-Objects.pdf

Wai, J. (2021). *incrementalPCA* [Computer software]. https://www.mathworks.com/matlabcentral/fileexchange/88872-incrementalpca)

Waidmann, E. N., Koyano, K. W., Hong, J. J., Russ, B. E., & Leopold, D. A. (2022). Local features drive identity responses in macaque anterior face patches. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-33240-w

Wallis, G. (1996). Using Spatio-temporal Correlations to Learn Invariant Object Recognition. *Neural Networks*, *9*(9), 1513–1519. https://doi.org/10.1016/S0893-6080(96)00041-X

Wallis, G. (2002). The role of object motion in forging long-term representations of objects. *Visual Cognition*, *9*(1–2), 233–247. https://doi.org/10.1080/13506280143000412

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, *98*(8), 4800–4804. https://doi.org/10.1073/pnas.071028598

Wallraven, C., Schwaninger, A., Schuhmacher, S., & Bülthoff, H. H. (2002). View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-extra-Ortho. In H. H. Bülthoff, C. Wallraven, S.-W. Lee, & T. A. Poggio (Eds.), *Biologically Motivated Computer Vision* (pp. 651–660). Springer. https://doi.org/10.1007/3-540-36181-2_65

Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical Imaging of Functional Organization in the Monkey Inferotemporal Cortex. *Science*, *272*(5268), 1665–1668. https://doi.org/10.1126/science.272.5268.1665

Wang, G., Tanifuji, M., & Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research*, *32*(1), 33–46. https://doi.org/10.1016/S0168-0102(98)00062-5

Wang, Y., Dong, X., Li, G., Dong, J., & Yu, H. (2022). Cascade Regression-Based Face Frontalization for Dynamic Facial Expression Analysis. *Cognitive Computation*, *14*(5), 1571–1584. https://doi.org/10.1007/s12559-021-09843-8

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. https://doi.org/10.3758/BF03202828

Watson, D. M., & Johnston, A. (2022). A PCA-Based Active Appearance Model for Characterising Modes of Spatiotemporal Variation in Dynamic Facial Behaviours. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.880548

Watson, T., Johnston, A., Hill, H., & Troje, N. (2005). Motion as a cue for viewpoint invariance. *Visual Cognition*, *12*(7), 1291–1308. https://doi.org/10.1080/13506280444000526

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(6982), Article 6982. https://doi.org/10.1038/nature02420

Webster, M. A., & MacLeod, D. I. A. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1702–1725. https://doi.org/10.1098/rstb.2010.0360

Weibert, K., & Andrews, T. (2016). The Relative Role of Viewpoint and Identity in the Neural Representation of Faces in Fusiform Gyrus. *Journal of Vision*, *16*(12), 715–715. https://doi.org/10.1167/16.12.715

Weibert, K., Flack, T. R., Young, A. W., & Andrews, T. J. (2018). Patterns of neural response in face regions are predicted by low-level image properties. *Cortex*, *103*, 199–210. https://doi.org/10.1016/j.cortex.2018.03.009

Weibert, K., Harris, R. J., Mitchell, A., Byrne, H., Young, A. W., & Andrews, T. J. (2016). An image-invariant neural response to familiar faces in the human medial temporal lobe. *Cortex*, *84*, 34–42. https://doi.org/10.1016/j.cortex.2016.08.014

Weiner, K. S., Golarai, G., Caspers, J., Chuapoco, M. R., Mohlberg, H., Zilles, K., Amunts, K., & Grill-Spector, K. (2014). The mid-fusiform sulcus: A landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. *NeuroImage*, *84*, 453–465. https://doi.org/10.1016/j.neuroimage.2013.08.068

Welling, L. L. M., Jones, B. C., Bestelmeyer, P. E. G., DeBruine, L. M., Little, A. C., & Conway, C. A. (2009). View-Contingent Aftereffects Suggest Joint Coding of Face Shape and View. *Perception*, *38*(1), 133–141. https://doi.org/10.1068/p5656

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

Wojciulik, E., Kanwisher, N., & Driver, J. (1998). Covert Visual Attention Modulates Face-Specific Activity in the Human Fusiform Gyrus: fMRI Study. *Journal of Neurophysiology*, *79*(3), 1574–1578. https://doi.org/10.1152/jn.1998.79.3.1574

Wright, A., & Barton, J. J. S. (2008). Viewpoint invariance in the discrimination of upright and inverted faces. *Vision Research*, *48*(25), 2545–2554. https://doi.org/10.1016/j.visres.2008.08.019

Wu, Y., & Ji, Q. (2019). Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision*, *127*(2), 115–142. https://doi.org/10.1007/s11263-018-1097-z

Xiao, N. G., Quinn, P. C., Ge, L., & Lee, K. (2012). Rigid facial motion influences featural, but not holistic, face processing. *Vision Research*, *57*, 26–34. https://doi.org/10.1016/j.visres.2012.01.015

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*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yang, H., Susilo, T., & Duchaine, B. (2016). The Anterior Temporal Face Area Contains Invariant Representations of Face Identity That Can Persist Despite the Loss of Right FFA and OFA. *Cerebral Cortex*, *26*(3), 1096–1107. https://doi.org/10.1093/cercor/bhu289

Yang, P.-L., & Beck, D. M. (2021). Does familiarity influence discrimination? Famous and Inverted Faces and Logos. *Journal of Vision*, *21*(9), 2001. https://doi.org/10.1167/jov.21.9.2001

Yang, P.-L., & Beck, D. M. (2022). Images that humans rate as highly representative of their category serve as better training for machine learning. *Journal of Vision*, *22*(14), 3311. https://doi.org/10.1167/jov.22.14.3311

Yang, P.-L., & Beck, D. M. (2023). Familiarity influences visual detection in a task that does not require explicit recognition. *Attention, Perception, & Psychophysics*, *85*(4), 1127–1149. https://doi.org/10.3758/s13414-023-02703-7

Yetter, M., Robert, S., Mammarella, G., Richmond, B., Eldridge, M. A. G., Ungerleider, L. G., & Yue, X. (2021). Curvilinear features are important for animate/inanimate categorization in macaques. *Journal of Vision*, *21*(4), 3. https://doi.org/10.1167/jov.21.4.3

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979. https://doi.org/10.1126/sciadv.aax5979

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Yin, X., & Liu, X. (2018). Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition. *IEEE Transactions on Image Processing*, *27*(2), 964–975. https://doi.org/10.1109/TIP.2017.2765830

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational Information in Face Perception. *Perception*, *16*(6), 747–759. https://doi.org/10.1068/p160747

Young, R., & Lesperance, R. (2001). The Gaussian Derivative model for spatial-temporal vision: II. Cortical data. *Spatial Vision*, *14*(3–4), 321–389. https://doi.org/10.1163/156856801753253591

Young, R., Lesperance, R., & Meyer, W. (2001). The Gaussian Derivative model for spatial-temporal vision—I. Cortical model. *Spatial Vision*, *14*, 261–319. https://doi.org/10.1163/156856801753253582

Yovel, G., & Freiwald, W. A. (2013). Face recognition systems in monkey and human: Are they the same thing? *F1000Prime Reports*, *5*, 10. https://doi.org/10.12703/P5-10

Yovel, G., Halsband, K., Pelleg, M., Farkash, N., Gal, B., & Goshen-Gottstein, Y. (2012). Can massive but passive exposure to faces contribute to face recognition abilities? *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 285–289. https://doi.org/10.1037/a0027077

Yovel, G., & Kanwisher, N. (2004). Face Perception: Domain Specific, Not Process Specific. *Neuron*, *44*(5), 889–898. https://doi.org/10.1016/j.neuron.2004.11.018

Yovel, G., & Kanwisher, N. (2005). The Neural Basis of the Behavioral Face-Inversion Effect. *Current Biology*, *15*(24), 2256–2262. https://doi.org/10.1016/j.cub.2005.10.072

Yue, X., Cassidy, B. S., Devaney, K. J., Holt, D. J., & Tootell, R. B. H. (2011). Lower-Level Stimulus Features Strongly Influence Responses in the Fusiform Face Area. *Cerebral Cortex*, *21*(1), 35–47. https://doi.org/10.1093/cercor/bhq050

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, *222*, 117295. https://doi.org/10.1016/j.neuroimage.2020.117295

Zachariou, V., Del Giacco, A. C., Ungerleider, L. G., & Yue, X. (2018). Bottom-up processing of curvilinear visual features is sufficient for animate/inanimate object categorization. *Journal of Vision*, *18*(12), 3. https://doi.org/10.1167/18.12.3

Zhang, H., Japee, S., Stacy, A., Flessert, M., & Ungerleider, L. G. (2020). Anterior superior temporal sulcus is specialized for non-rigid facial motion in both monkeys and humans. *NeuroImage*, *218*, 116878. https://doi.org/10.1016/j.neuroimage.2020.116878

Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020). Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 356–363. https://doi.org/10.1109/FG47880.2020.00134

Zhao, C., Seriès, P., Hancock, P. J. B., & Bednar, J. A. (2011). Similar neural adaptation mechanisms underlying face gender and tilt aftereffects. *Vision Research*, *51*(18), 2021–2030. https://doi.org/10.1016/j.visres.2011.07.014

Zhao, M., Bülthoff, H. H., & Bülthoff, I. (2016). A shape-based account for holistic face processing. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *42*(4), 584. https://doi.org/10.1037/xlm0000185

Zhou, L., Yang, A., Meng, M., & Zhou, K. (2022). Emerged human-like facial expression representation in a deep convolutional neural network. *Science Advances*, *8*(12), eabj4383. https://doi.org/10.1126/sciadv.abj4383

Zhou, Y., Liu, X., Feng, X., & Zhou, G. (2021). The constancy of the holistic processing of unfamiliar faces: Evidence from the study-test consistency effect and the within-person motion and viewpoint invariance. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-021-02255-8

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2013). *Deep Learning Identity-Preserving Face Space*. 113–120. https://www.cv-foundation.org/openaccess/content_iccv_2013/html/Zhu_Deep_Learning_Identity-Preserving_2013_ICCV_paper.html

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014a). Multi-View Perceptron: A Deep Model for Learning Face Identity and View Representations. *Advances in Neural Information Processing Systems*, *27*.

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014b). Recover Canonical-View Faces in the Wild with Deep Neural Networks. *arXiv*. https://doi.org/10.48550/arXiv.1404.3543

Zimmermann, F. G. S., & Eimer, M. (2013). Face learning and the emergence of view-independent face recognition: An event-related brain potential study. *Neuropsychologia*, *51*(7), 1320–1329. https://doi.org/10.1016/j.neuropsychologia.2013.03.028