Towards a unified account of face (and maybe object) processing

by

Cheston Y.-C. Tan

B.S., Electrical Engineering and Computer Science
University of California, Berkeley (2003)

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN NEUROSCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2012

Signature of Author ……..
Cheston Tan
Department of Brain and Cognitive Sciences
April 25, 2012

Certified by …….
Tomaso A. Poggio
Eugene McDermott Professor
Thesis Supervisor

Accepted by …….
Earl K. Miller
Picower Professor of Neuroscience
Director, Brain and Cognitive Sciences Graduate Program

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Towards a unified account of face (and maybe object) processing

by

Cheston Y.-C. Tan

Submitted to the Department of Brain and Cognitive Sciences
on April 25, 2012 in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Neuroscience

ABSTRACT

Faces are an important class of visual stimuli, and are thought to be processed differently from objects by the human visual system. Going beyond the false dichotomy of same versus different processing, it is more important to understand how exactly faces are processed similarly or differently from objects.

However, even by itself, face processing is poorly understood. Various aspects of face processing, such as holistic, configural, and face-space processing, are investigated in relative isolation, and the relationships between these are unclear. Furthermore, face processing is characteristically affected by various stimulus transformations such as inversion, contrast reversal and spatial frequency filtering, but how or why is unclear. Most importantly, we do not understand even the basic mechanisms of face processing.

We hypothesize that what makes face processing distinctive is the existence of large, coarse face templates. We test our hypothesis by modifying an existing model of object processing to utilize such templates, and find that our model can account for many face-related phenomena. Using small, fine face templates as a control, we find that our model displays object-like processing characteristics instead.

Overall, we believe that we may have made the first steps towards achieving a unified account of face processing. In addition, results from our control suggest that face and object processing share fundamental computational mechanisms. Coupled with recent advances in brain recording techniques, our results mean that face recognition could form the "tip of the spear" for attacking and solving the problem of visual recognition.

Thesis Supervisor: Tomaso A. Poggio
Title: Eugene McDermott Professor

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

“All models are wrong, but some are useful”

– George E. P. Box

"Robustness in the Strategy of Scientific Model Building"
in *Robustness in Statistics*. Edited by Launer RL, Wilkinson GN.
New York: Academic Press; 1979: 201-236.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

This thesis is dedicated to the memory of my father

TAN Chwee Huat, Ph.D.

11 April 1942 – 29 November 2011

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Acknowledgements

First and foremost, thanks to my advisor, Tommy Poggio. I can say with confidence that because of the fantastic match of intellectual interests and working style, Tommy is the best possible Ph.D. advisor I could have had. I also deeply appreciate the fact that Tommy allowed me to have an unbelievable amount of independence and latitude, while at the same time also providing generous access to many resources and opportunities.

Thanks to my thesis committee: Jim DiCarlo, Pawan Sinha and David Sheinberg. Each in their own way, they have variously been critical yet supportive, and have provided a wealth of insight, perspective and experience – not just for this thesis, but for the various other projects that are not included here.

Thanks to Nancy Kanwisher, who, although not a member of my thesis committee, has contributed as much as anyone else in shaping this thesis. Nancy was involved with this project from its inception, and has provided many supportive nudges to help it grow.

Thanks to Thomas Serre and Gabriel Kreiman for their patient guidance in my early years as a grad student and in various other projects, and also for being young faculty role models that I can look up to.

Thanks to various members and alumni of the CBCL family, for providing relaxation through foosball and pool, as well as stimulation through discussions and collaborations. A special mention of the people I've interacted most closely with: Ethan, Gadi, Hueihan, Joel, Jim, Kathleen and Sharat.

Thanks, not least, to family and friends, for providing emotional support and an ineffable sense of rootedness that ironically spans Singapore, Cambridge, Berkeley, Hong Kong, Lund and wherever they have been.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Table of Contents

Copyright notices

Chapter 1: Introduction

Faces are an important class of object stimuli, and are thought to be processed differently from non-face objects. However, faces and non-faces share at least some processing in common, e.g. the initial processing in area V1. So, going beyond the false dichotomy of same versus different processing, it is more important to understand how exactly faces are processed similarly or differently from non-faces. If face and non-face processing are radically different, then understanding one may not contribute much to understanding the other. On the other hand, if face and non-face processing share fundamental similarities, then understanding one may help with the other.

But why is this important? Faces are an interesting stimulus class in their own right, but they can also be viewed as a "model class". In biology, scientists often study "model organisms" such as C. Elegans, fruit flies, zebrafish or mice because these organisms have certain properties such as simplicity or short reproductive cycles that make them especially useful. After gaining some fundamental understanding from these model organisms, scientists then use that knowledge to better understand more and more complex organisms, hoping to eventually translate that understanding into medical treatments for humans and so on.

What makes faces a suitable candidate to be a "model class"? If we want to understand and ultimately replicate the amazing visual recognition abilities of humans, then it makes sense that a model class should be one that humans are good at recognizing, such as faces. We would also want the model class to be relatively well-defined and easily characterized. Faces fulfil these criteria, at least more so than many other classes such as fruits, clothes or animals. Finally, for practical purposes it would be greatly helpful if there existed a set of neurons that are spatially clustered and highly selective for the model class. Faces fit the bill perfectly.

With that broad motivation in mind, how well do we understand the processing of this model class? Not very well. Various aspects of face processing, such as holistic, configural, face-space and norm-based processing are investigated in relative isolation, and the detailed relationships between these are unclear. Furthermore, face processing is characteristically affected by various stimulus transformations such as inversion, contrast reversal and spatial frequency filtering, but how or why is unclear. Most importantly, we do not yet understand even the basic mechanisms underlying face processing.

There are two main messages in this thesis. The first is that we may be on our way to having a unified account of face processing. The second is that face processing and object processing may share striking similarities in underlying processing mechanisms. Coupled with recent advances in brain recording techniques, our results mean that face recognition could form the "tip of the spear" for attacking and solving the problem of visual recognition.

What do we mean by a unified account, and what has not been unified? We examine this next.

[Note: although strictly speaking, faces are themselves an object class (as opposed to non-object stimulus classes such as white noise, textures, etc.), throughout this thesis, we will use the terms "non-face objects", "objects" and "non-faces" interchangeably.]

1.1 Gaps in understanding

This thesis is a modeling study. If the purpose of scientific research is to gain a deeper understanding of the world we live in, how can models help to achieve that purpose? Empirical studies not only report about collected data, but also attempt to interpret that data. However, interpretation cannot exist in a vacuum, and relies on some model or theoretical framework. Empirical studies and modeling studies are thus complementary in the sense that they can work together to help us achieve an understanding that neither alone is capable of. We contend that the understanding of face processing is not advancing as quickly as it should, because of the lack of a unifying framework. We describe three major gaps in understanding currently faced by the field.

1.1.1 Multi-faceted face processing

There are multiple facets to face processing, and the relationships between these are still not clear. Under what is sometimes termed "configural processing", three varieties are distinguished: first-order, holistic, and second-order processing (Maurer et al. 2002). There is also another set of inter-related processing styles: face-space and norm-based coding. The relationships between all of these types of processing are still not well understood (McKone 2010).

Faces also seem to be different from objects in terms of their sensitivity to stimulus transformations such inversion, contrast reversal and spatial frequency filtering. How and why this is the case is not well understood.

1.1.2 From neurons to behavior

Another huge gap in understanding is in relating one level of description to another. Scientists have recorded from "face cells" in the macaque temporal lobe for decades (e.g. Gross et al. 1972), and these electrophysiological techniques have recently been given a radical boost through the advent of fMRI-targeting (Tsao et al. 2006). At the same time, behavioral studies of face processing in humans have proceeded in parallel for many decades (e.g. Yin 1969). In more recent years, these behavioral studies have been augmented by large-scale brain recording techniques such as fMRI and EEG.

However, there is still a yawning gap between each of these levels of description. It is still somewhat of a mystery how the responses of face-sensitive cells in the macaque temporal lobe relate mechanistically to human behavioral responses during face processing. How fMRI and EEG data relate to either of these is even more unclear, because there are multiple face-sensitive brain regions, and not all the findings from each region are consistent with behavior (Schiltz et al. 2010)

1.1.3 Faces and objects

It is clear that faces and objects are processed differently in some ways and similarly in others. What does this mean in terms of the underlying processing mechanisms? Are faces and objects similar because they share processing stages (e.g. V1) and are different because their processing stages diverge later? Or are face and object processing similar-yet-different because their processing mechanisms are variations of each other?

1.2 Hypothesis, results and contributions

In this section, we briefly describe our main hypothesis. We then provide a quick preview of the results and contributions; these will be described in more detail in Chapter 13 (and justified through the course of this thesis).

1.2.1 The "large coarse templates" hypothesis

Our main hypothesis is that the key difference between face and non-face processing lies in the existence of large, coarse face templates. The initial motivation and inspiration for this hypothesis is the Composite Face Effect (reviewed in Chapter 2) and holistic processing, thought by some to be the key characteristic that distinguishes face processing from object processing. This hypothesis/theory is more fully described in Section 13.2.

1.2.2 Preview of results and contributions

Our first result is that our model with large, coarse templates reproduces the Composite Face Effect (CFE), a "gold standard" behavioral marker of "holistic processing". With small, fine templates, the model does not show the CFE, consistent with our hypothesis. We then extend our investigation to inverted faces, and we find that our model can reconcile some of the conflicting behavioral results for inverted faces. Similarly, we find that our model can also reconcile some of the conflicting results stemming from two hotly debated versions of the CFE.

We then turn to the effect of three stimulus transformations on face processing: contrast reversal, spatial frequency filtering and inversion. For each of these stimulus transformations, we show that our model can once again reconcile some of the contradictory findings. Crucially, by using the same model with large, coarse templates without any modification (and using small, fine templates as a control), we demonstrate the link between these stimulus transformations and face-specific (holistic) processing.

Finally, we again use the same model and show that it can account for other kinds of processing (e.g. configural and face-space processing) thought by some to be face-specific, thereby linking these various aspects of face processing under our "large, coarse" framework.

At every step along the way, we provide a mechanistic, step-by-step account starting from the responses of single model units up to the level of "behavioral" responses by the model. We also show that a change in "processing style" (i.e. "face-like" versus "object-like"), even using faces

as stimuli, can account for the differences between behavioral results that have been found for faces and objects.

In short, we bridge each of the three gaps in understanding highlighted in Section 1.1. The contributions of our model and its implication are discussed in more detail in Chapter 13.

1.3 Thesis overview

We start off in Chapter 2 by doing an in-depth review of the Composite Face Effect (CFE), a "gold standard" marker of holistic face processing, because this is the starting point that motivates our "large, coarse template" hypothesis. In Chapter 3, we then review existing models of face processing to examine the current gaps in theoretical understanding. In Chapter 4, we review the HMAX model of object recognition, as this is the base model which we will modify to implement our "large, coarse template" hypothesis. We then proceed to verify our hypothesis in Chapter 5, by replicating the CFE. In Chapter 6, we extend our results by accounting for the CFE in inverted faces. We further extend our results in Chapter 7 by also accounting for an alternative CFE experimental design. In Chapters 8, 9, and 10, we switch gears to testing our model's sensitivity to stimulus transformations such as contrast reversal, spatial frequency filtering and inversion, respectively. In Chapter 11, we again switch gears, and attempt to show that our model can also account for face-space processing (and norm-based coding in particular). In Chapter 12, we show that our quantitative model and the widespread intuitive account of the CFE are compatible. Finally, in Chapter 13, we summarize our results by presenting a new, unified theory of face processing, discuss its implications and then list some predictions and avenues for future work.

1.4 What makes a model useful?

"All models are wrong, but some are useful" – George Box (1979)

The basic requirement of any model is that it should replicate whatever phenomenon it was designed for. It should also generalize, i.e. be applicable beyond that phenomenon. In our case, a good model of the Composite Face Effect should obviously replicate it, but also be able to account for the lack (or reduction) of it for inverted faces and non-faces. Beyond these basics, however, the usefulness of a model can be judged in various other ways. Throughout the course of this thesis, we will show that our model can do all of these things.

1.4.1 Performing simulations

One major way in which models can be useful is in performing simulations. Simulations allow scientists to test their hypotheses concretely by implementing them as quantitative models, rather than relying on their (sometimes incorrect) qualitative intuition. Because quantitative models have to be implemented in order for simulations to be run, this forces some implicit assumptions to be made explicit and mechanisms to be specified in detail. This helps to improve scientific rigor. In addition, practical considerations (such as length of experiments) impose certain

constraints on empirical studies; model simulations face fewer constraints, so they may be able to provide a more comprehensive picture.

1.4.2 Making sense of data

Another way in which models are useful is in making sense of data. Often, the empirical results can be unclear or even conflicting. This may simply reflect experimental noise or fluke results, but could also reflect the complexity of the underlying phenomena being studied. Models can help to potentially reconcile these conflicting results. Furthermore, empirical studies are rarely identical in their experimental design or procedure, and model simulations can help to bridge the differences between studies. Perhaps most importantly, like with the Standard Model of particle physics and the DNA/RNA molecular model of genetics, models can provide an overarching theoretical framework within which empirical results are interpreted.

1.4.3 Making predictions

Finally, models are also useful for making predictions (phenomena that "fall out" from the model, but were not the phenomenon that the model was designed to produce). Some of the "predictions" may have already been tested empirically, and thus might be termed "post-dictions" instead. Nonetheless, these are still important, since the model was not specifically designed to show these. True predictions are extremely useful in suggesting experiments to further our understanding.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 2: The Composite Face Effect (CFE)

Chapter abstract

In this chapter, we conduct a detailed and critical review of the Composite Face Effect (CFE), one of the "gold standard" behavioral measures of holistic face processing. It should be noted that is a standalone review, conducted independently of our modeling work. There are two main takeaways from this chapter. The first is that the CFE is likely to be a disproportionate/differential effect (i.e. quantitatively different for faces versus objects, rather than qualitatively different). The second is that despite the current controversy over the two different experimental versions of the CFE, they are, in theory, equally valid in tapping holistic processing. Conflicting results are likely to stem from confounds in specific aspects of experimental procedure and analysis, rather than experimental design per se.

Chapter contents

Chapter 2: The Composite Face Effect (CFE)

The Composite Face Effect (CFE) is widely acknowledged to be one of the "gold standard" behavioral markers of holistic/configural face processing (McKone & Robbins 2011, p.159). It is sometimes even presented as the "most convincing" demonstration of holism (Maurer et al. 2002, p.256; also McKone 2008, p.313). As such, a detailed understanding of the CFE may be intricately linked to that of holistic face processing. However, an in-depth and comprehensive review of the CFE does not exist (but see Richler et al. 2011a, suppl.). While reviews of face processing generally provide good coverage of the CFE, they often do so in the context of other issues of interest, such as development or expertise, and only in conjunction with other effects (Maurer et al. 2002, McKone et al. 2007, Tsao et al. 2010). In contrast, there are extensive reviews solely dedicated to the Face Inversion Effect (Valentine 1988, Rossion & Gauthier 2002, Rossion 2008). Therefore, the objective of this chapter is to provide an in-depth and critical examination of the CFE. We include detailed discussions of some controversies, and suggest ways to reconcile these. Ultimately, we examine if the CFE is indeed the best marker of holism.

We start by examining some limitations of other behavioral markers. The Face Inversion Effect (FIE), by itself, does not constitute evidence of face-specific processing (Maurer et al. 2002, p.258; McKone & Robbins 2011, p.161). This has been evident since the first study to demonstrate the FIE (Yin 1969; also see Valentine 1988, McKone et al. 2007), which explicitly stated that faces were disproportionately affected by inversion, in comparison to other objects that are customarily seen only in one orientation. The disproportionate (a.k.a. differential) nature of the FIE weakens the argument that face processing is qualitatively different. Furthermore, it is still not clear what the exact relationship between inversion and holism or second-order configuration might be.

The Whole-Part effect (WPE) is also a differential effect rather than an absolute one (Robbins & McKone 2007, p.50). Furthermore, the WPE may be due to a shift in criterion, rather than an improvement in sensitivity (Michel et al. 2006), or may be due to a contextual advantage not specific to faces (Gauthier & Tarr 2002, Leder & Carbon 2005). Perhaps most importantly, subjects are not given explicit instructions regarding which parts of the stimuli to attend to. Therefore, the WPE could simply reflect subjects choosing to take multiple features into account, so an advantage for wholes may be not surprising (Michel et al. 2006, Cheung et al. 2008). This does not accord with the general consensus that holistic processing for faces is automatic, not under decisional control (McKone & Robbins 2011, Richler et al. 2011d).

Does the CFE actually overcome the limitations faced by the FIE and WPE? To answer this question, we have to first describe the CFE in detail.

2.1 The "partial" experimental design

The CFE is tested using either the "partial" or "complete" experimental designs (see Fig. 2.1). The "partial" design is termed as such because it uses only a subset of the experimental conditions in the "complete" design. We describe this design first because it is simpler and more

widely used. (For an extensive listing of all “partial” and “complete” studies, see suppl. Table 2 of Richler et al. 2011a)

2.1.1 Identification and discrimination tasks

The CFE was first discovered using an identification task (Young et al. 1987). The top and bottom halves of faces belonging to different famous people were aligned together (forming “composites”), and subjects were tasked to identify the person shown in the top half. Reaction times (RTs) for this aligned condition were slower compared to when the top and bottom halves were misaligned by shifting the halves laterally.

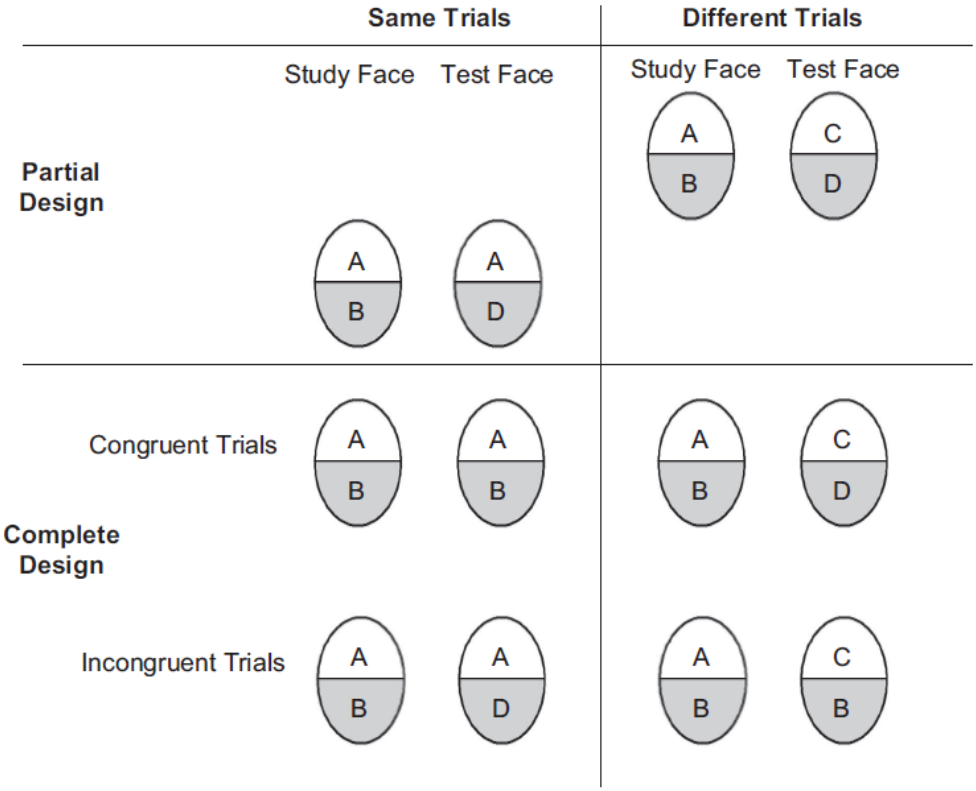


Figure 2.1. Trial types for the CFE “partial” and “complete” designs. Congruent trials are those in which the top and bottom halves are either both same or both different. Note that the “partial” design is a subset of the “complete” design. (Figure reproduced from Cheung et al. 2008. See p.16 for copyright notice)

Later, this experiment was extended to unfamiliar faces (Hole 1994) by using a discrimination (matching) task instead. Two composites are shown, either simultaneously or sequentially, in each trial. The top halves can be the same or different, while the bottom halves are always different. Subjects are instructed to ignore the bottom halves, and determine if the top halves of the two composites are the same or different. Importantly, because the bottom halves are always different, good performance on the “same” trials (which are the trials of interest) is contingent on ignoring the bottom halves as much as possible. Like with the identification task, performance on

the “same” trials is found to be better when the halves are misaligned instead of aligned. For this discrimination task, performance is defined as accuracy (i.e. hit rate, since only “same” trials are of interest), reaction time, or both.

In both the identification and discrimination versions of the experiment, the standard interpretation of the results is as follows. When the top and bottom halves are aligned, the composite faces are automatically processed holistically. Therefore, perception of the top halves is influenced by the bottom halves (despite subjects trying to ignore the bottom halves). In the case of identification, the identity of the top half takes longer to match, since the bottom half belongs to a different person. In the case of discrimination, the top halves – which are in fact identical – seem different, because the bottom halves are different. When the halves are misaligned, holistic processing is “disrupted”, reducing the influence of the bottom halves and therefore producing better performance. We call this the “misalignment effect”, to differentiate it from the “congruency effect” that is found using the “complete” paradigm (Section 2.2).

Note that in the discrimination task, the “different” trials are not analyzed because there is no clear prediction (at least by this intuitive account of the CFE; but see Section 12.4). Since the top halves are already different, the influence of the bottom halves could make them either “more different” or “less different”. Note that this logic is only applicable to the “partial” design, but has been incorrectly applied to the “complete” design also (see Section 2.2).

2.1.2 Robustness of the “partial” design

Since the CFE was first discovered in 1987, the effect has been replicated many times under a variety of experimental conditions, making it a highly robust effect (arguably as robust as the FIE). Multiple studies have used both the identification (Young et al. 1987, Carey & Diamond 1994, Khurana 2006, Singer & Sheinberg 2006, McKone 2008) and discrimination tasks (most other studies). The CFE has been found for RT (Young et al. 1987, Carey & Diamond 1994, Hole 1994, McKone 2008), hit rate (de Heering et al. 2007, Mondloch & Maurer 2008) and both (most other studies). The faces used have been famous (Young et al. 1987, Khurana et al. 2006, Singer & Sheinberg 2006), personally familiar (Carey & Diamond 1994), familiarized through training (Carey & Diamond 1994, McKone 2008) and novel (most other studies).

Presentation times have ranged from fairly short times of 80ms (Hole 1994) and 200ms (Le Grand et al. 2004, Mondloch & Maurer 2008), to unlimited time (Carey & Diamond 1994, de Heering et al. 2007, McKone 2008). For the discrimination task, the two composite faces have been presented simultaneously (Hole 1994, Hole et al. 1999, de Heering et al. 2007, Robbins & McKone 2007) or sequentially (most other studies). For sequential presentation, the faces have also been masked (Taubert & Alais 2009, Soria Bauser et al. 2011).

Furthermore, the CFE has been replicated under more demanding conditions such as image jittering (Robbins & McKone 2007, Rossion & Boremanse 2008), as well as size and brightness/contrast differences between two sequentially-presented composites (Robbins & McKone 2007) to prevent dependence on low-level cues. The half to be ignored can be either top or bottom (Young et al. 1987, Robbins & McKone 2007), and the CFE is independent of gaze behavior (de Heering et al. 2008)

Thus far, we have listed studies that have replicated the CFE under changes to variables that any robust effect would be expected to, e.g. task, presentation time and jitter. We now turn to studies that have found the CFE even for more high-level or drastic changes. These are nonetheless unsurprising results that are either predicted intuitively, or consistent with findings from other markers like the FIE. These may therefore be considered further evidence of robustness (this is subjective, however).

The CFE has been found for vertically-split halves (Hole 1994), horizontally-split but vertically-misaligned halves (Taubert & Alais 2009) and even for interior/exterior composites (Young et al. 1987, Singer & Sheinberg 2006). It is seen for in-plane rotations of up to 60 degrees (Mondloch & Maurer 2008, Rossion & Boremanse 2008), for frontal, 3/4 and profile views (McKone 2008), and even when the two composites within each trial are of different views (Hole et al. 1999). It has been demonstrated when misalignment only occurs for the second composite (Michel et al. 2006), when the alignment/misalignment is manipulated using the flash-lag effect (Khurana et al. 2006) and when the face "halves" that form composites are separated by 80ms of noise (Singer & Sheinberg 2006). Finally, the CFE is also seen for blurred or low spatial frequency (LSF) versions of faces (Goffaux & Rossion 2006, Cheung et al. 2008, Taubert & Alais 2011).

2.1.3 Results derived from the "partial" design

We now list a sampling of findings that have utilized the "partial" design to investigate holistic processing in faces – findings that are either surprising or at least not obviously predicted. Again, such characterizations are subjective.

In children, the CFE has been found at ages 10 (Carey & Diamond 1994), 6 (Carey & Diamond 1994) and even 4 (de Heering et al. 2007). This suggests that holistic processing is either present at birth or rapidly develops in infancy. However, the CFE is absent in patients with early visual deprivation – even after many years of subsequent visual experience (Le Grand et al. 2004), favoring the latter explanation. Furthermore, in adults, holistic processing was found for children's faces, and the duration of each subject's experience with children's faces was correlated with the difference between misalignment effect magnitude for adult and children faces (de Heering & Rossion 2008). This further supports the theory that holistic face processing is experience-dependent. (Note that this does not imply that experience alone is sufficient to produce holistic processing – either for faces or objects.)

Interestingly, the CFE is larger for same-race than other-race faces (Michel et al. 2006), although it is unclear if other-race faces are processed holistically (results regarding this were different for Caucasian and Asian subjects).

Although contrast reversal is detrimental for recognition and discrimination, three separate studies have found that contrast-reversed faces are nonetheless processed holistically (Hole et al. 1999, Calder & Jansen 2005, Taubert & Alais 2011), as are faces that are both blurred and contrast-reversed (Taubert & Alais 2011). (See Chapter 8 for an account of these seemingly contradictory phenomena)

Related to recent theories regarding the special status of horizontal information in face processing (Dakin & Watt 2009, Goffaux & Dakin 2010), the CFE is more tolerant to vertical than horizontal misalignment (Taubert & Alais 2009).

Finally, consistent with many other findings regarding hemispherical asymmetry, a larger CFE was found for left visual field (right hemisphere) presentation (Ramon & Rossion 2012). However, this was found only for the discrimination task, not the identification task.

2.2 The "complete" experimental design

As mentioned earlier, the "partial" design uses a subset of the conditions in the "complete" design (see Fig 2.1). In the "complete" design, the phenomenon demonstrating holism is the sensitivity (i.e. D') being higher in the congruent condition than in the incongruent condition, termed the "congruency effect".

The intuitive explanation of how the "congruency effect" relates to holistic processing is conceptually very similar to the explanation for the "partial" design. The logic is as follows (and best understood using Fig. 2.1 as a guide). In the incongruent trials, by definition, the bottom halves are opposite of the top halves. By this, we mean that if the top halves are the same, the bottom halves are different (and vice-versa). If the composites are processed holistically, then incongruent bottom halves exert detrimental influence on the top halves. Top halves that are identical seem different, while top halves that are different seem more similar. Therefore, holistic processing causes decreased sensitivity. In contrast, for the congruent trials, the top halves and the bottom halves are in agreement (by definition). It is debatable whether this makes top halves that are identical "more identical", or top halves that are different "more different". However, it is clear that compared to incongruent trials, congruent trials are less detrimentally influenced. Therefore, the D' sensitivity for the congruent trials should be greater than for the incongruent trials. If the halves are misaligned, then this difference in D' should decrease (or disappear). In the limit, when the halves are completely removed, there is no difference in D' between congruent and incongruent trials (since they are now in fact indistinguishable).

Importantly – and it is surprising that this has not been discussed elsewhere – it is not actually necessary to combine "same" and "different" trials into a D' measure to see the effect of holistic processing. The logic is identical to that above, but let us go through it slowly. Consider the "same" trials first. Comparing the congruent versus incongruent "same" trials, it should be obvious that holistic processing would cause the top halves in the incongruent trials to seem more different, compared to the congruent trials. This implies a lower accuracy (hit rate) for incongruent trials.

Now consider the "different" trials. Comparing the congruent versus incongruent "different" trials, it should also be intuitively obvious that holistic processing should make the top halves in the incongruent trials become more similar (relative to the congruent trials). This is regardless of whether B and D are more similar to each other than A and C, because B and D cannot be more similar than B and B. As a result, the accuracy (correct-rejection rate) for incongruent trials should be lower than for congruent trials. (In practice, because A and C are already clearly

different, the bottom halves may not be able to make them sufficiently similar, so as to be considered "same" – but that is an empirical matter)

Since for both "same" and "different" trials, the accuracies (hit rate and correct-rejection rate, respectively) are lower for incongruent than congruent trials, therefore it automatically follows that the D' for incongruent trials should be smaller.

Does this logic pan out empirically? Only two studies involving the "complete" design publish accuracies separately for "same" and "different" trials (Cheung et al. 2008, Gauthier et al. 2009). A detailed examination of the numbers (Cheung et al. 2008, Appendices B & C; Gauthier et al. 2009, Table 1) confirms that the trends are as predicted.

As we have seen, the intuitive logic for both the "partial" and "complete" designs is essentially identical. However, the "complete" design has been incorrectly criticized on the grounds that there is no clear prediction for the congruent "different" trials (McKone & Robbins 2007, 2011). By itself, that statement is correct, and is a valid argument against analyzing the "different" trials in the "partial" design. However, when considered in relation to the incongruent "different" trials, it is clear that a clear theoretical prediction can be made.

In the "partial" design, there were two task versions: identification and discrimination. Thus far, the "complete" design has primarily been used for discrimination, with the three exceptions being Singer & Sheinberg (2006), Richler et al. (2009b) and Cheung et al. (2011). The "complete" design for the identification task is depicted in Fig. 2.2. Whereas the "partial" design utilizes only the A/B composite, the "complete" design uses both A/A and A/B composites. Here, the A/A condition functions as a baseline with which to compare A/B performance, and the "congruency effect" refers to the difference in performance between A/A (congruent) and A/B (incongruent). This is very similar to the discrimination task, whereby the two congruent conditions can be construed as baselines for the two incongruent conditions.

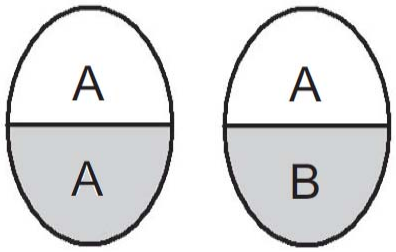


Figure 2.2. Trial types for the "complete" design identification task. The "partial" design uses only the A/B composite (right). The A/A composite is congruent, since both halves belong to the same person. The A/B composite is incongruent, since the halves belong to different people.

The identification task in the "partial" design relies on a comparison to the misaligned condition in order to establish holistic processing. However, despite the popularity of misalignment as a method to (ostensibly) "disrupt holism", it is still unclear exactly what the effect of misalignment is. By comparing A/A to A/B, misalignment is not strictly necessary to establish holism. (Without holistic processing, performance for A/A and A/B should in theory be identical) This may make the identification task an important complement to the discrimination task.

2.2.1 Robustness of the "complete" design

While the "complete" design is generally less commonly used than the "partial" design, it is also fairly robust, and the congruency effect has been replicated under different conditions. Presentation of the two composites in each trail can be simultaneous (Goffaux 2009) or sequential (most other studies). The presentation times can as short as 200ms/200ms for the first/second composite (Richler et al. 2011a), 800ms/50ms (Richler et al. 2009c, Richler et al. 2011c), and up to 3 seconds (Goffaux 2009). The half to be ignored can be either top or bottom (Richler et al. 2008a, 2008b, 2009, 2011b, 2011c, Goffaux 2009, Gauthier et al. 2009). The halves can be jittered (Goffaux 2009) or masked (Richler et al. 2008a, 2008b, 2009, 2011b, c, Gauthier et al. 2009, Cheung & Gauthier 2010). The halves in the two composites can be both aligned (Cheung et al. 2008, Richler et al. 2008b, Gauthier et al. 2009), or only one aligned (Richler et al. 2008a, 2008b, 2011a, 2011b, 2011c). Finally, the congruency effect is also found for low spatial frequency (LSF) faces (Cheung et al. 2008, Goffaux 2009).

2.2.2 Results derived from the "complete" design

Using the "complete" design, it was found that holistic face processing has a significant decisional component (Richler et al. 2008a), but neither perceptual nor decisional accounts alone is sufficient (Richler et al. 2008b). The congruency effect can be independent of top-down beliefs and also the ratio of "same" versus "different" trials (Richler et al. 2011b). A face memory load reduced the congruency effect, whereas an object memory load did not (Cheung & Gauthier 2010).

A congruency effect has been found for high spatial frequency (HSF) faces to different degrees (significant: Cheung et al. 2008; marginal: Goffaux 2009). Inverted faces are also processed holistically, but require longer presentations times for holism to manifest (Richler et al. 2011c).

In normal subjects, the magnitude of holistic processing (congruency x alignment interaction) is correlated with face recognition ability (Richler et al. 2011a). In individuals with Autism spectrum disorders (ASDs), the congruency effect is less affected by misalignment, compared to control subjects (Gauthier et al. 2009).

Many "complete" design studies investigated the issue of expertise for non-faces (e.g. Gauthier et al. 2003, Hsiao & Cottrell 2009, Wong et al. 2009a, b, Bukach et al. 2010, Wong & Gauthier 2010). The reader is referred to Gauthier et al. 2010 for a comprehensive review.

2.3 Inverted faces

The face inversion effect (FIE) (Yin 1969) is the most established and well-known face-related phenomena. Its acceptance is such that inversion is commonly used as a control in face studies. If an effect exists for faces, but not for inverted faces (or at least significantly reduced), then the effect is claimed to be face-specific.

The logic is not watertight, however. Inverted faces are sometimes purported to be processed "like objects" (Haxby et al. 1999, Richler et al. 2011c), but multiple studies have found that inverted faces also activate face-selective areas significantly – albeit less then upright faces (Kanwisher et al. 1998, Haxby et al. 1999, Yovel & Kanwisher 2005). Furthermore, the FIE is a differential effect. In other words, inversion affects both objects and faces, and the key signature is that it affects face significantly more than objects. Therefore, if an effect is merely reduced for inverted faces, then that is at most weak evidence for face specificity, since the effect of inversion on object-processing mechanisms might account for the reduction. If an effect is absent for inverted faces, then that is stronger evidence for face-specificity, but it is still not conclusive. In particular, if an effect is weak, then the effect of inversion on object-processing mechanisms alone may be sufficient to account for the absence of the effect.

Earlier, we discussed the limitations of the FIE and WPE, including the fact that they are differential effects. The CFE is therefore sometimes preferred to the FIE and WPE as a marker for face-specific processing (and holistic processing in particular). Here, we critically examine whether the CFE is actually free of this limitation, in relation to inverted faces (below) and non-faces (Section 2.4). We tackle this issue separately for the "partial" and "complete" designs.

2.3.1 Inverted faces: "partial" design

Many studies utilizing the "partial" design either do not include an inverted condition (e.g. Le Grand et al. 2004, Michel et al. 2006, de Heering et al. 2007, etc.), while others use inversion as the primary image manipulation rather than misalignment (Hole 1994, Hole et al. 1999, Singer & Sheinberg 2006). Of the studies that include inverted faces as a control for the misalignment effect in upright faces, there is no unanimous agreement.

The studies that do not find a misalignment effect for inverted faces outnumber those that do. The following studies do not find a misalignment effect: Young et al. 1987 (Table 2), Carey & Diamond 1994 (Figs. 2, 4, and 5), Robbins & McKone 2007 (Fig. 7), McKone 2008 (Fig. 3), Mondloch & Maurer 2008 (Figs. 3 and 5) and Soria Bauser et al. 2011 (Fig. 3). In contrast, two studies found a misalignment effect for inverted faces (albeit significantly reduced): Goffaux & Rossion 2006 (Fig. 5) and Rossion & Boremanse 2008 (Figs. 3 and 4).

How can these findings be reconciled? Before dismissing the results of Goffaux & Rossion (2006) and Rossion & Boremanse (2008) as statistical flukes or outliers, the differences between all of these studies deserve closer scrutiny. Perhaps the most salient factor is whether upright and inverted faces were shown in different blocks. If solely inverted faces were shown in a particular block, subjects could have used part-based strategies to perform the task, possibly leading to the absence of a misalignment effect. This, by itself, does not refute the claim that inverted faces are

processed non-holistically, but that is not the claim that researchers are actually trying to make. Rather, it is the claim that given identical conditions and task strategy (to the extent possible), upright faces are processed holistically, while inverted faces are not (or are processed less holistically). To demonstrate this, upright and inverted faces should be randomly intermixed. This cannot fully guarantee that systematic strategy differences are absent, but it is much better than blocked trials.

Empirically, among the studies that do not find a misalignment effect for inverted faces, most have upright and inverted faces in separate blocks. The trials in Soria Bauser et al. 2011 are also blocked, but it is unclear what is being blocked. Experiment 1 of Mondloch and Maurer (2008) is apparently the sole study to randomly intermix orientations. However, aligned/misaligned trials were blocked, and it is unclear what effect this may have.

On the contrary, Goffaux and Rossion (2006) had different sets of subjects for the upright and inverted conditions (meaning that subjects for the inverted condition might have used part-based strategies), and yet they nonetheless found a misalignment effect for inverted faces. If orientations were intermixed, this effect could potentially have been even stronger. Rossion & Boremanse (2008) do not explicitly describe what is blocked in their study. However, they had 5 blocks but 7 orientation conditions, so we presume that orientation was randomized within each block. This study also found a misalignment effect for inverted faces.

Altogether, it is possible that a misalignment effect was not found in many studies primarily because inverted and upright faces were blocked. Other factors may also contribute, since the studies were fairly diverse in their experimental conditions (e.g. task, presentation times, etc.) Another potentially important factor is the statistical analyses that were performed – but this has not been investigated thoroughly. Standard ANOVA (e.g. Carey & Diamond 1994) assumes that data in the different conditions are drawn from independent samples, and determines if the means in each condition differ significantly. Inter-subject variability (e.g. individual subjects performing very well or very poorly overall) tends to make the means in the various conditions more similar and the variances larger, therefore potentially masking the differences. To compensate for inter-subject variability, paired two-sample t-tests (with Bonferroni-correction) should be used.

2.3.2 Inverted faces: "complete" design

For the "complete" design, only two studies thus far have investigated the congruency effect for inverted faces. Similar to the "partial" design, the results are mixed. No congruency effect was found by Goffaux (2009, see Fig. 3), while congruency effects were found by Richler et al. (2011c, see Figs. 2 and 5). These two studies had numerous differences, e.g. simultaneous versus sequential presentation, so it is difficult to discern the cause of the conflicting results.

2.3.3 Inverted faces: summary

In sum, considering both the "partial" and "complete" designs, inverted faces have been found to be processed holistically or non-holistically in different studies. Unless the results of the former studies can be explained away, the common denominator is that inversion reduces holistic

processing, sometimes to the extent of completely disrupting it (or at least making it undetectable).

This has important implications for the CFE and the study of face processing. One limitation of the FIE and WPE is that they are differential rather than absolute effects. The effect of inversion on the CFE suggests that it may also be a differential effect after all. If this is true, then among the three "gold standard" markers of holistic face processing, all are differential effects. This would then imply that we do not currently have a behavioral marker that qualitatively differentiates holistic from non-holistic processing, and therefore we also cannot conclusively claim that faces are qualitatively different from objects in terms of holistic processing.

However, there is one other avenue for potentially redeeming the special status of the CFE as an absolute effect: non-faces. We examine this next.

2.4 Non-faces

If it is important to establish that the CFE is face-specific (i.e. exists for faces but not non-faces), then why not test this directly? Somewhat surprisingly, this has not be done extensively (for either "partial" or "complete" design), especially in comparison to the FIE. For the FIE, comparing inversion of faces and objects is the standard control used to establish face-specificity. Perhaps due to the widespread acceptance of inversion as disproportionately affecting faces, studies using the CFE tend to rely on inverted faces rather than objects as control stimuli. This is truer for the "partial" design, whereby inverted faces are used to demonstrate the absence (or reduction) of the misalignment effect. For the "complete" design, studies tend to use misaligned (rather than inverted) faces to demonstrate the absence (or reduction) of the congruency effect. There may also be practical considerations for avoiding objects as control stimuli – extra stimuli do not need to be collected if inverted faces are used. But perhaps most importantly, inverted faces are very similar to faces in terms of physical characteristics. Apart from differences in phase, the spatial frequency content is exactly the same in upright and inverted versions of an image. This provides certain advantages to using inverted stimuli as controls.

Nonetheless, given that we still do not fully understand how inverted faces are processed, it is useful and important to examine the CFE for non-face objects. We separately consider studies using the "partial" and "complete" designs. In the studies that included object experts and novices, we only considered the novices. We deliberately avoid the issue of holistic processing in object experts, and refer the reader to the existing literature (e.g. Gauthier et al. 2010, McKone & Robbins 2011)

2.4.1 Non-faces: "partial" design

We are aware of only three studies involving non-face objects using the "partial" design. No misalignment effect was found for dogs (Robbins & McKone 2007, Fig. 7). Similarly, no misalignment effect was found for bodies, either without heads or with faceless heads (Soria Bauser et al. 2011, Figs. 3 and 4). With frontal views of cars, Macchi Cassia et al. (2008) overall

found no misalignment effect. Interestingly, however, for short (200ms) presentation times, a misalignment effect may have been found for both hit rate and reaction time (Macchi Cassia et al. 2008, Fig. 3). Statistical significance was not reported, but visually, the SEM error bars were above zero (suggesting significance). Furthermore, for reaction time, the effect for faces and cars was not significantly different ($p=0.728$). For longer (600ms) presentation times, however, a misalignment effect was clearly absent.

2.4.2 Non-faces: "complete" design

In stark contrast to the studies using the "partial" design, most studies using the "complete" design found holistic processing for objects in novices. Using artificial novel stimuli, Wong et al. 2009a found a congruency effect for "Ziggerins" (see Fig. 5), while Richler et al. (2009a) did so for "Greebles" (Fig. 3). Similarly, a congruency effect was found for cars (Gauthier et al. 2003, Table 1; Bukach et al. 2010, Figs. 4 & 5), Chinese characters (Hsiao & Cottrell 2009, Figs. 6 and 7), and musical notes (Wong & Gauthier 2010, Figs. 2 and 3). However, Gauthier & Tarr (2002) did not find a congruency effect for "Greebles"; it is unclear why different results were obtained from Richler et al. (2009a).

Apart from the congruency effect, we also looked at the interaction between congruency and alignment (see Section 2.5.2 for reason) wherever possible. Richler et al. (2009a) found no interaction between congruency and alignment for "Greebles", meaning that the congruency effect was not significantly different for aligned or misaligned composites. In contrast, Bukach et al. (2010) found a strong main effect of misalignment on congruency effect (Fig. 4; $p<0.001$ overall, including experts). Hsiao & Cottrell (2009) found a marginally significant interaction for Chinese characters ($p<0.05$); the congruency effect was larger for aligned than misaligned characters. Wong et al. (2009a) did not explicitly analyze the statistical significance, but judging visually from Fig. 5b, there is potentially an interaction between congruency and alignment. Overall, the results are somewhat mixed for (congruency x alignment) interaction for object novices.

Interestingly, it seems that studies using the "partial" design tested more naturalistic objects such as bodies, dogs and cars (all of which can be seen in daily life), while studies using the "complete" design generally favored more artificial stimuli that are novel ("Ziggerins" and "Greebles") or not often seen (Chinese characters and musical notes). It remains to be seen whether this is a causal factor contributing to the starkly different results arising from the two paradigms.

2.4.3 Non-faces: summary

The "partial" and "complete" designs seem to result in generally opposite findings regarding holism in non-faces. (Note, however, that neither design had unanimous results.) These generally opposing results for non-faces may have significantly intensified the debate on which design is the correct (or better) one, a debate that was initially sparked by the opposite findings regarding the "expertise hypothesis" (Gauthier & Bukach 2007, Robbins & McKone 2007, McKone & Robbins 2011). We now critically compare the two designs in detail.

2.5 Comparing the "partial" and "complete" designs

In comparing the designs, the most basic question is whether they are consistent with holistic processing. The answer in both cases is yes. As described earlier, holistic processing makes certain predictions (misalignment effect and congruency effects) that have borne out empirically. The next question is in the reverse direction: do these effects necessarily imply holistic processing, or can there be alternative explanations?

2.5.1 Alternative explanations

For the "partial" design, Gauthier and colleagues have argued that biases have not been ruled out as alternative explanations for the misalignment effect (Richler et al. 2011c). In the "partial" design, "same" trials are also incongruent, while "different" trials are congruent. Does this confound matter? Empirically, Gauthier and colleagues found that misalignment causes a shift in bias towards a "same" response (e.g. Fig. 5B of Cheung et al. 2008, but see Fig. 3 of Richler et al. 2011c). This implies that the misalignment effect could, in theory, be simply due to this shift in bias, rather than disruption of holism. Moreover, a shift in bias (specifically, a congruency x alignment interaction for bias) is correlated with the misalignment effect across subjects (Richler et al. 2011b, Fig. 3). However, since the misalignment effect only uses "same" (incongruent) trials, it is unclear why the bias for congruent trials should be included for this correlation. Without the congruent trials, the correlation between bias and misalignment effect could be absent or much weaker (as hinted at by Figs. 3 and 4 of Richler et al. 2011c). Overall, the "bias-shift" explanation for the misalignment effect cannot be ruled out, but it is not conclusively proven yet. Importantly, there could be some underlying factor that causes both the bias-shift and the misalignment effect.

For the "complete" design, there is an alternative explanation in terms of "response competition". In Section 2.2, we described the logic behind the "complete" design in terms of what we call the "influence" account – for each composite, the bottom half influences perception of the upper half. Alternatively, one could also think in terms of the responses suggested by the top halves versus that suggested by the bottom halves. For example, for incongruent-same trials, the top halves suggest the response "same", while the bottom halves suggest "different", leading to lower accuracy. This "response competition" account is in some sense like the Stroop effect, and may not be face-specific. The key difference between the "influence" and "response competition" accounts is whether one compares composites or halves. The "influence" account combines the halves first, then compares composites. The "response competition" account compares the halves separately. The "response competition" account is less realistic, especially when the composites are presented sequentially. Since subjects are instructed to ignore the bottom halves, why would they subconsciously keep track of the bottom halves (separately from the top halves), compare the bottom halves and then allow that comparison to affect their behavioral response? It would be more parsimonious to suggest that separately for each composite, the two halves are perceptually integrated to some extent, i.e. the "influence" account. Nonetheless, the "response competition" account cannot be fully ruled out.

In addition, empirically, congruency effects can be induced contextually (Richler et al. 2009a), by mechanisms unknown. Furthermore, as discussed in Section 2.5.5, for the "complete" design, the first composite is often fully attended by subjects. This may lead to artificial inducement of seemingly "holistic" processing, a form of "holism" that could be Navon-like (Navon 1977), rather than the automatic holism that faces are thought to evoke. This is perhaps an explanation as to why "holism" is seen for objects – especially if measured by the congruency effect alone, but possibly even for a (congruency x alignment) interaction (see next Section).

This brings us to the question of what the right measure of holism should be, for both the "partial" and "complete" designs.

2.5.2 What should the right measure be?

Richler and colleagues (Richler et al. 2009a) found congruency effects for "Greebles" in novices. To account for this, they suggest that congruency effects can be contextually induced, and that the congruency effect by itself may not be a good measure. Instead, "*the comparisons of congruent and incongruent trials with a baseline and the interaction between congruency and alignment may be more indicative of holistic processing driven by the stimulus, as opposed to… context of the task"* (Richler et al. 2009a, p.538). In practice, they have more commonly used the (congruency x alignment) interaction (e.g. Richler et al. 2011a), rather than a baseline comparison. If it is indeed true that the congruency effect alone is not such a good measure, then studies that rely on this (e.g. Gauthier et al. 2003, Bukach et al. 2006, Richler et al. 2009c) must be interpreted cautiously. One problem with the (congruency x alignment) is that there is no sound theoretical basis for why it better reflects holism than congruency alone. The misaligned condition can be likened to a baseline, but this serves a practical function, not a theoretical one.

In addition, it is unclear if the signature of stimulus-driven, face-like holistic processing should simply be a (congruency x alignment) interaction, or the abolishment of a congruency effect for misaligned (or inverted) stimuli. Interestingly, congruency effects are abolished for misaligned (Cheung et al. 2008) or inverted (Goffaux 2009) faces when subjects are compelled to ignore one half for <u>both</u> faces. Conversely, when subjects are compelled to attend to the whole of the first face, congruency effects are mostly still found despite misalignment (Richler et al. 2008a, 2008b, Richler et al. 2011b, 2011c). This issue is examined in more detail in Section 2.5.5.

The issue of the right measure also applies to the "partial" design. Many studies (e.g. Le Grand et al. 2004, Michel et al. 2006, de Heering et al. 2007) have relied solely on a misalignment effect to draw conclusions about holism for children, other-race faces, etc., without including inverted faces or some other control. A (misalignment x inversion) interaction may be a more robust measure for the "partial" design; this is analogous to the (congruency x alignment) interaction for the "complete" design. One important reason is that the hit rates are often not equalized (either in the aligned or misaligned condition) when comparing populations, such as children versus adults. Without equalization, one cannot quantitatively compare different populations to determine if holistic processing is weaker or stronger. Otherwise, for example, de Heering et al. (2007) might have concluded that holistic processing is stronger in 4, 5, and 6 year-olds than in adults (see Fig. 4), which is counter-intuitive (though not necessarily wrong). Since equalization is not easy to achieve in practice, inverted faces or objects can act as controls. This issue is equally important

for accounting for individual differences when quantifying holism in individual subjects (Richler et al. 2011a, Wang et al. 2012). Note, however, that even if a (misalignment x inversion) interaction is used, the question of whether the misalignment effect should be abolished or reduced for inverted faces still remains (Section 2.3.1).

Asides from issues of what the right measure of holism should be, there are other criticisms of the "partial" and "complete" designs that deserve closer scrutiny.

2.5.3 Criticisms of "partial" design

The main argument against the "partial" design is the issue of biases, discussed earlier as an alternative explanation for the misalignment effect (also see Gauthier & Bukach 2007, Richler et al. 2011a, b). This issue deserves further investigation.

Considering the fact that the misalignment effect is abolished or significantly reduced for inverted faces and non-faces, it is hard to argue that generic biases are the primary cause of the misalignment effect (McKone & Robbins 2007). If the biases are specific to upright faces, then perhaps they are linked to holistic processing rather than other arbitrary factors. As Gauthier and colleagues themselves point out (Gauthier & Bukach 2007, p.326), the existence of these biases need not negate the claim that misalignment effects are caused by holism, as the causes and mechanisms for these biases are still unknown (Cheung et al. 2008, p.1333). One possible explanation is that holism is the underlying cause for both the biases and the misalignment effect.

Furthermore, the random intermixing of aligned and misaligned trials (e.g. Robbins & McKone 2007, etc.) strongly reduces the possibility that strategy-related biases that differ for aligned versus misaligned trials are the main cause of the misalignment effect.

Finally, the misalignment effect has been consistently found for the identification task also (e.g. Young et al. 1987, Carey & Diamond 1994). It is unclear how biases could account for the misalignment effects found this way (McKone & Robbins 2007).

2.5.4 Criticisms of "complete" design

The main argument against the "complete" design is that it does not measure "face-like" holism (McKone & Robbins 2011). It is argued that the "complete" design "*weakens the definition of holistic processing from… some form of very strong perceptual integration… to merely any failure of selective attention*", and that the congruency effect "*merely shows that competition for attentional resources from the to-be-ignored half is stronger when subjects are experts with the object class*" (McKone & Robbins 2011, p.163). This is identical to the "response competition" account described in Section 2.5.1.

The definition of holism as perceptual integration is indeed stronger than the definition as failure of selective attention. Perceptual integration leads to failure of selective attention, but the latter does not necessarily imply the former, as there multiple reasons for failure of selective attention.

Crucially, however, this criticism of the "complete" design is equally applicable to the "partial" design. Perception cannot be directly measured, and has to be inferred from behavior. Qualitative definitions aside, do the experimental designs actually differ in what they measure? No arguments have been put forth as to how the pattern of behavior analyzed in the "partial" design is able to measure perceptual integration better than the "complete" design can. As a matter of fact, the "same" trials in the "partial" design are incongruent, and the "response competition" (failure of selective attention) account is also applicable. As discussed in Section 2.2, the logic underlying both designs is essentially identical. If anything, the fact that the "partial" design is strictly a subset of the "complete" design means that the latter has the potential to be a more robust and sensitive measure, if analyzed correctly. This leads to a related issue.

Because of the way congruency is defined (i.e. whether there is agreement between the responses suggested by the top and bottom halves), the "complete" design is understandably seen by some as measuring the effect of competition by the two halves for attentional resources. This interpretation is not wrong – but it is also a valid interpretation of the "partial" design. Rather, it may be helpful to also separately analyze "same" and "different" trials. Lumping the two together to produce D' has its advantages, but puts the focus on the notion on "congruency" (and thus competition between the halves). By looking at just the "same" trials (for instance), the focus is now on whether the bottom halves make the top halves "more same" or "less same" (and thus invoking the notion of perceptual integration). The same logic applies to the "different" trials. Importantly, separate analysis of "same" and "different" trials acts as additional validation that the predictions made by holistic processing do actually occur. Combining trials by calculating D' may obscure potential failures of holistic processing or experimental artifacts.

A secondary argument against the "complete" design is that the D' measure includes the "different" trials, for which there is no theoretical prediction (McKone & Robbins 2007). This argument is incorrect, as discussed in Section 2.2.

In any case, an important point has been raised. Neither the "partial" or "complete" design has been shown to be able to dissociate perceptual integration from "merely" failure of selective attention. It is unclear if these can be dissociated behaviorally (electrophysiological recordings or brain imaging may be more suitable). Nonetheless, this issue deserves more thought.

2.5.5 Confounds in comparing "partial" and "complete" designs

We have argued that the logic underlying both experimental designs is essentially identical. Why, then, do the designs sometimes seem to produce starkly different results (e.g. for non-faces)? So far, criticisms have been focused on the designs themselves. If one design can be proven "wrong", then the results derived from it can then be ignored. However, these criticisms may have neglected to examine other differences between studies that produce conflicting results. We examine the two most salient ones.

The first is whether subjects are compelled to fully attend to the first composite. In the majority of studies employing the "complete" design, subjects are forced to fully attend to the first composite, because they have not yet been cued which half to ignore. This cueing happens during the intervening blank, or even simultaneously with the second composite. Out of the 23

"complete" design studies listed in Suppl. Table 2 of Richler (2011a), only 20 were for the CFE per se (McKone & Robbins 2007, p.334). Of those 20, eight studies pre-cued subjects to ignore a specific half: Gauthier et al. 2003, Cheung et al. 2008, Gauthier et al. 2009, Goffaux 2009, Hsiao & Cottrell 2009, Cheung & Gauthier 2010, Todorov et al. 2010 and Richler 2011a. In contrast, to the best of our knowledge, all "partial" design studies pre-cued subjects to ignore a particular half (almost always the bottom).

If subjects fully attend to the first composite, how does this affect things? One might suspect that having to compare the second composite to a memory representation of a fully attended first composite may somehow induce the second composite to also be more fully attended. Furthermore, in some studies, the attentional cue (a bracket, shown either above or below the face) is only revealed concurrently with the second composite (e.g. Richler et al 2008a, Cheung & Gauthier 2010, Wong & Gauthier 2010, expt. 1 of Richler et al. 2011c). Subjects are thus ironically forced to attend to the entire stimulus in order to find out which half they have to ignore. Together, these may artificially induce what might seem to be "holistic" processing. Importantly, this is more like Navon-type global processing, rather than the automatic and unavoidable holistic processing that is thought to be face-specific.

Consistent with this hypothesis, congruency effects are abolished for misaligned (Cheung et al. 2008, Gauthier et al. 2009) or inverted (Goffaux 2009) faces when subjects are compelled to ignore one half for both faces. Conversely, when subjects are compelled to attend to the whole of the first face, congruency effects are mostly still found despite misalignment (Richler et al. 2008a, 2008b, Richler et al. 2011b, 2011c).

Can fully attending the first composite account for the counter-intuitive findings of congruency effects for objects in novices? Of the six studies that found such effects (see Section 2.4.2), four studies compelled subjects to fully attend to the first composite; these are consistent with our hypothesis. We now examine the other two studies more closely. Hsiao & Cottrell (2009) studied Chinese characters using simultaneous presentation of both composites. Because the composites were presented vertically (one above the other) and simultaneously, subjects may have been unable to properly ignore the appropriate halves in their haste to look at both composites, which were presented for 600ms. Sequential or simultaneous-horizontal presentations might have negated this confound. Gauthier et al. (2003) interleaved face and car composites and found congruency effects for both. The interleaving may have contextually induced holistic processing (see Gao et al. 2011 and Fig. 5 of Richler et al. 2009a) for the cars. (Note: there is no evidence that the "partial" design is any more or any less resistant to contextual inducement than the "complete" design). In addition, car expertise (as defined by an independent measure) fell along a continuum, and "novices" were simply defined as those below the median. Thus, it is not clear that congruency effects were actually found for car "novices" in the proper sense. Overall, there is no clear evidence refuting the hypothesis that a procedural confound (fully attending to the first composite or not), rather than the "partial" or "complete" design per se, is responsible for congruency effects in object novices. Therefore, while further studies are needed to verify this hypothesis, the "partial" and "complete" designs may potentially agree on the issue of holism in novices after all.

Another important procedural confound between "partial" and "complete" studies is whether the first composite is always aligned (even for "misaligned" trials) or not. Among the "complete" design face studies that have a misaligned condition, the first composite was always aligned in the following four studies: Richler et al. 2008a, and Richler et al. 2011a, b, c. Conversely, two "complete" design face studies had both composites misaligned: Cheung et al. 2008 and Gauthier et al. 2009. In contrast, among the many more "partial" design studies, as far as we are aware, only Expt. 1 of de Heering et al. (2007) had the first composite aligned in the "misaligned" trials.

Interestingly, of the two studies that had both composites misaligned, both found that misalignment completely removed the congruency effect, not just reduced it (Cheung et al. 2008, Fig. 5; Gauthier et al. 2009, Fig. 4). In contrast, in the studies that had only the second composite misaligned, misalignment often only reduced the congruency effect, rather than removed it (Richler et al. 2008a, Fig6; Richler et al. 2011b, Figs. 3 & 6; Richler et al. 2011c, Figs. 2 & 5). This may not be surprising. If misalignment disrupts holism, then misalignment of only one composite may disrupt holism less than misalignment of both composites. If the vast majority of the "partial" design studies misalignment both composites, while many "complete" design studies do not, then this factor may be an important confound in comparing these studies.

Why do some "complete" design studies (especially the more recent ones) always have the first composite aligned? Richler et al. (2009a) reported that congruency effects were found in "Greeble" novices when the first composite was misaligned, while no congruency effects were found when the first composite was aligned. These results were interpreted as showing that holism can be induced even in novel objects simply by having the first composite misaligned. As a result, subsequent studies by Gauthier and colleagues have had the first composites always aligned, in order to avoid such effects. (While we do not disagree with their interpretation, we believe that the crucial factor was the fact that the first composite was fully attended. This, in turn, was what allowed the misalignment of the first composite to induce a congruency effect, by necessitating a larger attentional window)

Of special interest are the results of two studies that include all four combinations (first composite aligned/misaligned crossed with second composite aligned/misaligned) for faces (Richler et al. 2008b) and "Greebles" (Richler et al. 2009a). Subjects in both studies always fully perceived the first composite. For faces, when the second composite was misaligned, whether the first face was misaligned or not generally had small effects, especially on the magnitude of the congruency effect (see Figs. 4 and 6 of Richler et al. 2008b). Trial order was randomized in this study. For "Greebles", the same result was found, but only when trial types were randomly intermixed (Richler et al. 2009a, Fig. 3). When alignment of the first composite was a between-subjects variable, this had a significant effect (Richler et al. 2009a, Fig 2A). In other words, consistent with what we have discussed earlier, the alignment/misalignment of the first composite may affect processing strategy.

In summary, differences in results from "complete" versus "partial" design studies cannot be solely attributed to the design per se, because of two confounds in experimental procedure (full perception and misalignment of the first composite). We have shown, prima facie, that these confounds may in fact be the reason for conflicting results. Specifically, these confounds may

act, separately or in concert, to give the appearance of a congruency effect for misaligned faces or non-faces. Further studies are needed to confirm this hypothesis.

2.5.6 Studies analyzing both "partial" and "complete" designs

In theory, any study employing the "complete" design can also analyze only the subset of trials found in the "partial" design, and compare the results of both designs. In practice, only in recent years has this been done. We look at each of the four studies that do so.

Cheung et al. (2008) examined the CFE for spatial frequency (SF) filtered faces. Analyzing the "partial" design trials, they replicated the finding of Goffaux & Rossion (2006) that low spatial frequency (LSF) faces elicit a larger misalignment effect than high spatial frequency (HSF) faces. However, using the "complete" design, they found that both LSF and HSF faces elicited congruency effects and (congruency x alignment) interactions of similar magnitude. Like most "partial" design studies, Cheung et al. (2008) randomly intermixed trials, always aligned/misaligned both composites, and subjects knew beforehand which half to ignore. It would seem, therefore, that they genuinely found that different experimental designs can produce different results. However, Goffaux (2009) also used the "complete" design, but found larger congruency effects and (congruency x inversion) interactions for LSF than HSF faces. Thus, while Goffaux (2009) did not compare "partial" and "complete" designs, and therefore does not directly refute the claim that experimental design can influence findings, the issue is still not fully resolved.

Richler et al. (2011a) use the "complete" design and found that the (congruency x alignment) interaction for D' and RT were correlated with face recognition performance across subjects. On the other hand, the alignment effect was not found to be significantly correlated. It would then appear that the "complete" design produces a more robust measure of holism. However, it must be noted that the "complete" design analyzed 4 times the number of trials than did the "partial" design ("same"/"diff" x congruent/incongruent trials, compared to just "same" trials). Furthermore, the average correlation for the "partial" design was 0.143 (compared to 0.311 for the "complete" design), so the trend is in the right direction. If the number of trials had been matched, and/or an (alignment x inversion) interaction had been used instead, it is possible that the "partial" design correlations would have been significant too. (Inversion for the misalignment effect acts as a baseline control, analogous to misalignment for the congruency effect)

Richler et al. (2011b) tries to discount the "partial" design by showing that the results are influenced by extraneous factors such as the perceived or actual same/diff trial mixture (e.g. 25% same 75% different versus the opposite case). The "complete" design was found to be resistant to such factors. Richler et al. (2011c) compared upright and inverted faces. It was found that the (congruency x alignment) interaction did no differ for upright and inverted faces. However, the misalignment effect did. However, in both studies, the same issues that were found for Richler et al. (2011a) also exist: 4 times the number of trials are used for "complete" design analyses, and the aid of a baseline control (misaligned condition) for the "complete" design.

In short, even in studies that directly compared "complete" and "partial" designs, it is still not clear if the design differences per se (rather than number of trials or use of baseline controls) contribute to the seeming weakness of the "partial" design.

2.5.7 The way forward

In comparing the "partial" and "complete" designs, we have touched on several different experimental and theoretical issues, such as the right metric to use, that have yet to be resolved. It is important that these issues be resolved, if we are to fully understand holism and face processing. Here, based on what we have discussed earlier, we summarize our views on some of these issues.

We view both "partial" and "complete" designs as equally valid, in the sense that holistic processing predicts both the misalignment and congruency effects, and the logic behind these is essentially identical. In practice, one may be better than the other, not due to the design per se, but due to practical considerations. For example, the (congruency x alignment) interaction uses the misaligned condition as a baseline control, something which is missing in the misalignment effect. A (misalignment x inversion) interaction may be a better metric, just like the (congruency x alignment) interaction seems to be a better metric than the congruency effect alone.

The issue of biases is an important one, even if we believe that it does not necessarily invalidate the "partial" design. There could be a common causal factor that produces both a bias shift and the misalignment effect. Nonetheless, when comparing conditions that are in separate blocks, and especially when comparing different populations, this is an important issue to take into account.

As such, and since there does not seem to be any advantage to the "partial" design, we advocate using the "complete" design, with the following important details. We believe that the subjects should know beforehand which half to ignore, and that both halves should be either aligned or misaligned. Furthermore, to minimize effects of strategy, all trials should be intermixed. Analysis-wise, omnibus ANOVAs are not sufficient; within-subjects measures (adjusted for multiple comparisons) are more sensitive, since they sidestep inter-subject variability. Finally, before combining hit rate and false-alarm rate to produce D', these should be separately verified to each show a congruency effect (or at least not a negative congruency effect). One secondary advantage of the "complete" design is efficiency: all the trials are used in the analysis, unlike the "partial" design, which discards "different" trials.

There's little doubt that any measure of holism can be affected by non-stimulus factors, e.g. Gao et al. 2011, Richler et al. 2009a, Richler et al. 2011a, b. What's important is to minimize this. How should we go about isolating "face-like", i.e. automatic and unavoidable, holism? By trying to prevent it as much as possible, e.g. clear demarcation of top/bottom halves, short presentations to avoid eye movements, side-by-side (rather than top-bottom) placement for simultaneous presentation, pre-cueing so that only the to-be-attended halves ever need to be processed, and misaligning by shifting only the to-be-ignored halves (rather than both halves). Altogether, the experiment should try to aid the subjects in perceiving only the halves to be attended. Thus, any measured holism would be intrinsic rather than incidental to the experimental procedure.

These measures serve to maximize the automatic, perceptual nature of the task. In other words, subjects should be able to maximally focus their attention on perceiving the to-be-attended half. Under such circumstances, the "amount of holistic processing" that is measured can be thought to be purely perceptual. Other factors, such as context and priming can indeed affect holistic processing, and we want to measure the holistic processing that is independent of these factors, which may not be face-specific.

One remaining concern is that of using the (congruency x alignment) interaction as a metric, instead of just the congruency effect. The misaligned condition acts as a baseline control. For example, if for whatever reason the D' is artificially boosted, because both aligned and misaligned trials should be equally affected, subtracting away the misaligned performance negates this artificial boost. However, there is no good theoretical reason why the congruency effect should not be an adequate measure of holism. Further thought should be given to this issue.

2.6 Qualitative or quantitative

We began this chapter by stating that the FIE and WPE are differential effects, meaning that they cannot prove that face processing is qualitatively different. This chapter has found that the CFE may also be a differential effect after all.

But does it really matter if face processing is "merely" quantitatively different? The "special" status of face processing is not necessarily solely contingent on it being qualitatively different. For example, having a dedicated brain area for face processing does not necessarily mean that this brain area performs qualitatively different processing. Furthermore, faces can be processed both as parts and as wholes, and inverted faces activate both face and object selective areas. Finally, the definition of face-ness is not necessarily a binary one – stimuli can vary continuously in terms of how face-like they are. Overall, there is really no reason that face processing should be expected to be qualitatively different from object processing.

In fact, even if only faces are processed holistically, holism itself is not an all-or-none phenomenon, as evidenced from the CFE magnitude at various angles of rotation (Mondloch & Maurer 2008, Rossion & Boremanse 2008). The amount of misalignment also affects the magnitude of the misalignment effect (Richler et al. 2008a, Fig. 6; Taubert & Alais 2009, Table 1). In other words, it seems fruitless to search for a behavioral measure of holism that is absolute rather than differential, because holism itself is not an absolute, all-or-none phenomenon.

Ultimately however, it may be a mistaken endeavor to attempt to infer qualitativeness/quantitativeness from the magnitude of behavioral measures. Just because something is qualitatively different does not necessarily mean that it will produce a large measurable difference. Conversely, quantitative differences can also produce small or large measurable differences. Take, for example, two V1 simple cells tuned to orthogonal orientations. Most people would consider the cells to be quantitatively different (differing only in the angle of tuning), but the cells would respond very differently to most stimuli, especially oriented gratings.

On the contrary, take a V1 simple cell and a V1 complex cell, both tuned to the vertical orientation. Assuming that the complex cell pools over several different simple cells, one might consider the complex cell to be qualitatively different from the simple cells. Under most circumstances, the responses of the complex cell and the simple cells would generally be similar, with some differences only in terms of position tolerance. Therefore, qualitative and quantitative are defined by the underlying mechanisms, not by the size of the measurable effects.

2.7 Explanatory gaps

Thus far, we have highlighted several issues that should be more closely examined (and hopefully resolved) in future work. Among these are the differential versus absolute nature of the CFE, the CFE for inverted faces and non-faces, as well as the "partial" versus "complete" designs. Apart from these, there are some other ways in which our lack of understanding about holism is notably glaring.

2.7.1 From neurons to behavior

It is still not at all understood how holism arises mechanistically from the activity of single neurons. While neurophysiological studies have found hints of holistic processing in single neurons (Kobatake & Tanaka 1994, Freiwald et al. 2009), it is not clear how this gives rise to behavioral measures of holism, such as the CFE. More importantly, little is known about how holism in single neurons comes about.

2.7.2 Relation to other phenomena

The relationships between holism and various other major aspects of face processing are still not well understood. For example, inversion is commonly thought to "disrupt holism", but what exactly that means is unclear. Do inverted faces activate a population of neurons distinct from those activated by upright faces, and why are those neurons not holistic? Or is the same population of neurons activated, but simply to a lesser degree?

Similarly, "holistic" and "configural" processing are sometimes lumped together, but it is unclear why they should be (Maurer et al 2002). Processing a face "as a whole" is not necessarily the same thing as calculating the second-order distances between face parts. Another major gap in understanding is the relationship between "holistic/configural" processing and face-space processing (see Chapter 11), as discussed in several studies (Carey & Diamond 1994, McKone 2009a).

2.8 Perceptual integration versus selective attention

In their criticism of the "complete" design (see Section 2.5.4), McKone & Robbins (2011) have raised an important point. What should be considered "face-like" holistic processing, and what should not? McKone & Robbins (2011, p.163) define it as "very strong perceptual integration", criticizing the definition as "failure of selective attention" to be too general.

However, do these different definitions actually make different predictions, and are the current behavioral measures able to discriminate these definitions? In general, investigation into these issues is lacking. In McKone (2008), which uses the "partial" design identification task, only top halves were named, so that interference from the bottom halves could not possibly be due to Stroop-like semantic effects. Richler et al. (2009b) argued that the CFE is not due to "response interference", unlike the Stroop effect. More emphasis needs to be given to resolving this issue, however.

2.9 Chapter summary

In this chapter, we have conducted a detailed and critical review of the Composite Face Effect (CFE). It is considered to be one of the "gold standard" behavioral measures of holistic face processing, sometimes even the best measure. Our review has shown that the CFE, like other measures of holistic processing, may be just a differential effect after all.

One major issue in the literature is disagreement over whether the "partial" or "complete" design is better. Our review suggests that both may have equal theoretical validity, and conflicting results in the literature stem from confounds in experimental procedure, rather than the experimental design per se.

Overall, the CFE is still far from being fully understood. Does it measure "perceptual integration" or "failure of selective attention"? How exactly does inversion or misalignment affect holism, as measured by the CFE? How does the CFE relate to other measures or aspects of face processing? These questions remain to be answered.

Chapter 3: Models of Face Processing

Chapter abstract

The main objective of this chapter is to examine existing models of face processing to see where they fall short in explaining the empirical behavioral and electrophysiological data on face processing. As such, we focus on models from Neuroscience and Psychology, rather than Computer Vision. The chapter is organized by issue/phenomenon (e.g. the Face Inversion Effect, Composite Face Effect) and the models pertaining to each issue are compared critically. Overall, we find three main shortcomings of existing models. Firstly, few models provide a mechanistic, step-by-step account of how the relevant phenomena come about. Secondly, many models focus on accounting for face processing phenomena, but neglect to account for the lack of the same phenomena for objects and inverted faces. Third and most importantly, there is no existing model that provides a unified account of the multiple aspects of face processing examined here.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 3: Models of Face Processing

This chapter is a critical review that compares and contrasts existing models of visual face processing. In particular, we point out the deficits of each, and also the disagreements among different models. The emphasis is not on comprehensiveness. Instead, we focus on face detection and identification, and give particular emphasis to processing characteristics that are thought to be face-specific (e.g. holistic and configural processing). We generally favor quantitative models, but here we have included qualitative models for their theoretical insights. We deliberately exclude models of size, position and view invariance. These invariances are clearly important, but they are not specific to faces.

We also deliberately neglect models relating to expression and gaze. While these aspects of face processing are interesting and important, they are thought to be separate from processing of identity (Bruce & Young 1986, Haxby et al. 2000, but see Calder & Young 2005). Furthermore, expression and gaze processing clearly cannot precede some sort of face detection (or pseudo-detection) stage. We wish to focus on this first stage of face-specific processing (which may simultaneously perform identification, etc.), under the assumption that understanding this stage will be the key to unraveling the rest of face processing.

We begin our review with models of holism, which is the key issue that this thesis tackles.

3.1 Holism

In this section, we examine models that purport to explain holism, or at least replicate signatures of holistic processing. The main purpose here is to home in on the key aspect of each model that is supposed to produce holism. When this key aspect is not explicitly shown or stated by the authors, we attempt to decipher their work in order to extract it.

Since holism is still not well-defined or generally agreed upon (McKone & Yovel 2009, McKone & Robbins 2011), we restrict ourselves to what we think is possibly the most important aspect of holism – the ability of one portion of the face to influence the processing of another portion.

Examining a range of models, we find that they achieve this aspect of holism (shortened to just "holism" for the remainder of this section) generally via three mechanisms: "feature linkage", "feature combination" and "unitary representation". We examine each of these in turn.

The main idea behind feature linkage is to form dependencies between discrete face features, thereby allowing a change in one feature to affect the processing of others. The "fiducial point" model of Biederman & Kalocsai (1997) and the "ratio-template" model of Sinha (2002) rely on nameable face parts, whereas the "key point" model of McKone & Yovel (2009) and the model of Schwaninger and colleagues (Wallraven et al. 2005, Schwaninger et al. 2009) use salient or key points in an image without regard to what those points are. The dependencies can take the form of distances between features (Schwaninger et al. 2009, McKone & Yovel 2009), feature overlap (Rossion & Gauthier 2002), contrast ratios (Sinha 2002), or deformation costs

(Biederman & Kalocsai 1997). One major limitation of these models is that processing of inverted faces and non-faces is not well accounted for. For example, in theory, the distance between face parts can be calculated from the image just as easily for inverted as for upright images – so why do humans show the Face Inversion Effect (FIE)? For non-faces, there is generally little more than brief justification as to why such feature linkage mechanisms couldn't work or don't exist.

Models that take the "feature combination" approach (e.g. Turk & Pentland 1991, Cottrell et al. 2002, Rossion & Gauthier 2002, Jiang et al. 2006) generally come from a more neuroscience-oriented perspective, whereby neurons in higher-level visual areas are tuned to specific combinations of inputs from preceding visual areas. Hence, a change in a single local feature can change the response of a higher-level feature neuron, but that local feature cannot be pinpointed as the cause of the change just by looking at the neuron's response alone. If perception and behavior primarily utilize higher-level features, then the local features are "bound together" and cannot be processed separately. The tuning functions that combine local features can be linear, e.g. linear projections (Turk & Pentland 1991) or non-linear, e.g. gaussian radial basis functions (Jiang et al. 2006).

There are similarities between the "feature linkage" and "feature combination" approaches. The dependencies in "feature linkage" tend to be pair-wise and between a relatively small number of parts. The tuning to combinations of inputs in the "feature combination" approach can be viewed as creating dependencies between many features simultaneously, i.e. like "feature linkage", but on a larger scale.

However, models following the "feature combination" approach tend to be more biologically plausible and rely on simpler mechanisms. Moreover, the FIE has a straightforward account from these models: humans see many more upright than inverted faces, hence there are many neurons tuned to upright faces. Inverted faces elicit weak responses from such neurons, which may result in poor recognition for inverted faces. Overall, the "feature combination" models are more parsimonious.

The third approach to holism is via "unitary representation". Models following this approach tend to be qualitative and tend to adopt the traditions of behavioral psychology, i.e. they posit abstract psychological constructs to account for behavioral findings. Holism is attributed to "processing as a whole", or "a single, global representation", or "lack of decomposition into parts", or other numerous variations thereof. We coin the term "unitary representation" for the lack of a better summary description. Qualitative models following this approach generally state that upright faces are "processed as wholes", whereas inverted faces and non-faces are not, essentially just re-describing what is already known (or thought to be true). Models using this approach account for holism by appealing to the unitary nature of the representation.

Overall, all these models seem to have one thing in common that supposedly gives rise to holism: they bank on perception and behavior primarily utilizing representations from which the contributions of individual local inputs are not easily isolated. Note, however, that there is nothing necessarily singular, global or even "whole" about such a description of holism. These

representations can be distributed (e.g. Jiang et al. 2006), and they can each represent just a portion of the face (e.g. Sinha 2002).

We have analyzed these models in terms of the mechanisms that purportedly produce holism. However, not all of these models have actually been shown to demonstrate signatures of holistic processing. In the next two sections, we examine in more detail those models that claim to be able to show two such signatures of holism: the Face Inversion Effect (FIE) and the Composite Face Effect (CFE).

3.2 Face Inversion Effect (FIE)

The Face Inversion Effect (FIE) is a well-established phenomenon (see Chapter 10) dating all the way back to 1969 (Yin 1969). Surprisingly, not many models actually attempt to explicitly account for it. One possible reason is that it may seem obvious – any model can posit a bias for upright faces due to the fact that humans see way more upright than inverted faces, thereby easily accounting for the FIE.

However, upon closer examination, properly accounting for the FIE is in fact crucial to the veracity of any model. The FIE is actually the differential FIE: faces are more affected by inversion than non-faces are (Yin 1969, Valentine 1988, McKone et al. 2007). Therefore, simply arguing that upright faces are seen more often than inverted faces is insufficient, because humans do not see upright and inverted houses, cars, etc. in equal proportion either. We now turn to the few models that attempt to explain the FIE, and see if they also account for the differential nature of the FIE.

Jiang and colleagues (Jiang et al. 2006) use a shape-based model to quantitatively account for not just the FIE itself, but also the patterns of human performance for featural and configural changes in upright and inverted images (also see Section 3.7). Crucially, they also investigate the factors in their model that give rise to the FIE, finding that it is the tight tuning of their face units to upright faces that is the key. Accordingly, they explain the differential FIE as stemming from tight tuning for face units and broader tuning for car (or other non-face) units. Why the difference in tuning? They posit that expertise with faces causes this. Interestingly, this accords with the finding that dog experts have a large inversion effect (Carey & Diamond 1986, but see McKone et al. 2007 p.10). To summarize, Jiang and colleagues argue that the FIE is caused by exposure to more upright than inverted faces, and the differential FIE is caused by more expertise with faces than non-faces. Importantly, they specify that these differences are manifested in neuronal tuning properties.

Zhang & Cottrell (2004, 2006) similarly tackle the effect of inversion on featural and configural changes. Like Jiang et al. (2006), they reproduce the FIE. However, they do not examine in detail the mechanisms that give rise to the FIE, nor attempt to account for the differential nature of the FIE.

Like the two models above, McKone & Yovel (2009) attempt to account for the effect of inversion on featural versus configural changes, but qualitatively instead of quantitatively. They

hypothesize that "processing of upright faces encompasses not only spacing between the major features, but the detailed shape of those features as well" (p.788). As with similar models (that calculate distances between features), they do not detail the mechanisms of how the FIE occurs. We presume that it boils down to the exposure argument: upright faces are more commonly seen. As noted above, this is somewhat unsatisfactory, because in terms of image properties, the distances can just as easily be calculated for inverted faces. McKone & Yovel also do not attempt to account for the differential FIE.

Valentine (1991) attempts to account for the FIE, especially in relation to the distinctiveness effect (distinctive faces are better-remembered than typical faces). However, the explanation is simply that inverted faces elicit larger errors, without explanation why or how. Furthermore, the differential aspect of the FIE is not tackled.

Overall, thus far only one model (Jiang et al. 2006) has given a mechanistic account of the differential FIE, which is attributed to broader tuning width in object-tuned units, compared to face-tuned units.

3.3 Composite Face Effect (CFE)

As discussed in detail in the previous chapter, the Composite Face Effect (CFE) is one of the "gold standard" behavioral tasks used to gauge holistic processing.

Tsao and colleagues (Tsao & Livingstone 2008, Tsao et al. 2010) propose qualitatively that the CFE is the result of aligned faces being "obligatorily detected as a whole, but misaligned faces and cars are not" (Tsao & Livingstone 2008, Fig. 5 caption). One major problem with this explanation is that it does not give a good account of which stimuli get detected as faces or not. The magnitude of misalignment in Figure 5 of Tsao & Livingstone (2008) is fairly small, and the misaligned composite is still obviously very much like a face, yet the detector misses it (by their definition). Yet, many studies showing the CFE used aligned composites with distinct gaps between the halves (e.g. Goffaux & Rossion 2006, de Heering et al. 2007, Cheung et al. 2008, Rossion & Boremanse 2008) – should those be detected as faces if this slightly misaligned composite is not? If the entire left side of a face is occluded, should it be detected as a face? (It would seem to match a regular face far less than a misaligned composite) Furthermore, there is empirical evidence of the CFE for contrast-reversed (Hole et al. 1999, Taubert & Alais 2011), rotated (Mondloch & Maurer 2008, Rossion & Boremanse 2008) and even inverted (Rossion & Boremanse 2008, Richler et al. 2011c) faces, directly contradicting the proposal that "the filters… used by this detector stage require an upright, positive contrast face" (Tsao & Livingstone 2008, p.425). Lastly, it is not clear why faces should be processed "as a whole" to begin with. The authors suggest that detection allows for specialized neural machinery to be activated, in order to perform face-specific processing such as identification or gaze-tracking. However, none of this necessitates detection "as a whole" (detection can happen in various ways). On the contrary, certain tasks such as gaze-tracking or expression recognition would seem to be better off using fine-grained, part-specific, "non-holistic processing" (Zhang & Cottrell 2005), while it is still unclear whether identification is aided by "holistic processing" or not (Tsao & Livingstone 2008, Ullman et al. 2002).

Riesenhuber & Wolff (2009) state that their shape-based model (Riesenhuber & Poggio 1999, Jiang et al. 2009) "appears to also hold promise to account for some aspects of the CFE", briefly noting that their face-tuned model units respond poorly to misaligned composites. Their S2 templates are very small – covering only 2x2 C1 units (Jiang et al. 2006, p.169) – while their C2 units are invariant to position changes. Therefore, only a small fraction of the 256 C2 units should be affected by misalignment (those whose S2 templates happen to match the middle region of the face). In other words, their face-tuned units respond poorly to misaligned composites because of sharp tuning, whereby changes in a small number of C2 inputs has a large effect on face-tuned units. This is the same account that they propose for the FIE. Since object-tuned units have broader tuning, they are less affected by misalignment.

The model of Biederman & Kalocsai (1997) utilizes Gabor-jets (collections of V1-like responses for multiple orientations and scales) as the basic representation for faces, while also representing the spatial relationships between these jets, either as a deformable grid or as a graph of landmark points. The authors hypothesize that for misaligned composites, since the landmark points in the lower halves are not in their expected locations, therefore the jet values corresponding to these landmarks are not used, eliminating their influence on the top halves. This qualitative explanation is plausible, but their model is sensitive to contrast reversal, whereas it has been found that contrast-reversed faces still show the CFE (Hole et al. 1999, Taubert & Alais 2011; also see Chapter 8). Furthermore, the model does not specify how inverted faces are processed, so it is unclear whether a misalignment effect is predicted for inverted faces or not.

Unlike the model Biederman & Kalocsai (1997), the model of Cottrell and colleagues was actually used to concretely demonstrate the CFE, in both the "partial" (Cottrell et al. 2002) and "complete" (Richler et al. 2007) CFE designs (see Chapter 2). The model combined the ubiquitous V1-like Gabor-jet representation with Principal Components Analysis or PCA (inspired by eigenfaces) on the entire set of jets to produce a face representation that is holistic. However, little was done to advance the understanding of the nature of holism, especially in terms of giving a step-by-step account of the mechanisms that produce the CFE. In particular, it is not clear if it is PCA per se that is essential for the CFE, if any method that combines the entire set of jets will suffice, or if simply feeding the set of jets to the classifier (without combination) is enough. More generally, like some of the other models, this model has not been used to account for the CFE with respect to spatial frequency filtering, contrast reversal and inversion. Finally, no hypothesis is put forward to explain the absence of a composite effect for non-faces.

Overall, like with the FIE, only one model has actually concretely demonstrated the CFE (in this case, the model of Cottrell et al. 2002 and Richler et al. 2007). However, no detailed mechanistic account was proposed. Crucially, it was not shown that the CFE was absent (or reduced) for inverted faces and non-faces, thus missing essential controls.

3.4 Inverted faces and non-faces

Here, instead of focusing on specific behavioral effects (i.e. FIE and CFE), we attempt to compare and contrast the various models in terms of the way they account for how and why

inverted faces and non-faces are processed differently from upright faces. Note, however, that many models do not explicitly include such explanations, while others are simply unable to do so. Also note that we simply state these accounts without comment as to which are more plausible or whether empirical evidence favors one over another.

There are generally three accounts of the "specialness" of upright faces: stimulus characteristics, task demands, and exposure/expertise. These are not mutually exclusive. Also, some models have different accounts for explaining upright/inverted and face/non-face differences.

Stimulus characteristics are generally used to explain the face/non-face distinction, rather than the upright/inverted distinction. Faces are a fairly uniform class of stimuli, compared to other non-face classes. The major face parts are found in a standard general configuration, with the differences between individual faces stemming from the appearance of parts and the fine details of the configuration. As such, a face-detection stage may exist in order to determine if further face-specific processing should occur (Tsao & Livingstone 2008). Alternatively, for the task of identification, faces vary in terms of general configuration less than non-faces, so face processing requires mechanisms that are highly sensitive to fine configural changes (Biederman & Kalocsai 1997, Schwaninger et al. 2009). Either way, faces require processing that is specific to the class, and such processing will not be applied to non-faces.

The "task demands" account of the face/non-face distinction stems from the idea that faces need to be recognized at the individual level much more often than non-faces are. The computational demands of face identification may act in concert with stimulus characteristics (Biederman & Kalocsai 1997) or spatial frequency factors (Dailey & Cottrell 1997, 1999, Zhang & Cottrell 2005) to result in holistic representations and processing.

Finally, the exposure and expertise accounts can be used to explain both face/non-face and upright/inverted differences. Humans see faces more than any other single class of objects, and we also see upright faces more than inverted faces. As such, face processing units could become tightly tuned, while non-face units are more broadly tuned (Jiang et al. 2006). This account may also need to rely on the "task demands" account, since expertise at detection (rather than identification) may not necessarily require tight tuning. This task-specific expertise account is favored by Zhang & Cottrell (2006). Most (possibly all) models rely on the exposure account to explain the upright/inverted differences, although many do so only implicitly. Note that apart from amount of exposure per se, there are also other reasons for humans being face experts (e.g. social demands and/or genetics), so exposure and expertise are not synonymous.

3.5 Detection versus identification

One important issue over which models disagree is which face processing stage (or task) gives rise to holism and holistic representations.

Tsao and colleagues (Tsao & Livingstone 2008, Tsao et al. 2010) explicitly proposed that "holistic face processing could be explained by the existence of an obligatory detection stage" (p.488 of Tsao et al. 2010). Similarly, the detection process seems to be the key to holism in the

work of Sinha (2002), since identification is not discussed, and the binary contrast-ratio representation seems to be unsuitable for the fine discriminations required for identification.

In direct contrast, the work of Cottrell & colleagues specifically argue in favor of identification. Dailey & Cottrell (1997, 1999) show that face identification leads to a specialized (but not clear if holistic) face processing area, whereas face detection does not. Using a related model, Zhang & Cottrell (2005) show that large ("holistic") image patches are more informative of identity than small image patches.

Biederman & Kalocsai (1997) do not discuss face detection but state explicitly that "individuation of faces… requires specification of the fine metric variation in a holistic representation of a facial surface… Such a representation will provide evidence for many of the phenomena associated with faces, such as holistic effects…" (p.1218 of Biederman & Kalocsai 1997).

Implicit bias towards identification is shown by several models. However, these focus on the interaction between inversion and featural/configural changes, and are thus not necessarily linked to holism. Riesenhuber and colleagues (Jiang et al. 2006) deal with same/different discrimination rather than identification per se (although these are related). In theory, their shaped-based model could be used for face detection. However, since the "inversion effect is based on tightly tuned model units" (p.163), such tight tuning may be unsuitable for face detection, while broader tuning removes the inversion effect. Likewise, Rossion's work proposing the qualitative "perceptual field hypothesis" (Rossion 2009) also makes no mention of detection. McKone and Yovel (2009) focus on identification, arguing qualitatively that there exists "a truly holistic representation of upright faces that integrates all details of the shape-related information for an individual face" (p.795), encompassing "not only spacing between the major features, but the detailed shape" too (p.788). Schwaninger and colleagues (Wallraven et al. 2005, Schwaninger et al. 2009) also see identification as the key to holism, as their face identification units "integrate featural and configural information to [form] holistic representations" (p.1436 of Schwaninger et al. 2009).

Overall, while more models seem to favor identification as the key to holism, some are agnostic about detection, rather than specifically arguing for identification. As such, there is no clear consensus on this issue. Is there some way to reconcile these two views? Interestingly, the eigenface representation (similar to the Gabor-PCA representation used by Cottrell and colleagues, who explicitly argue in favor of identification) can be used for both detection and identification (Turk & Pentland 1991). Furthermore, upon closer inspection, the size of features that have been optimized for either identification (Zhang & Cottrell 2005) or detection (Ullman et al. 2002) turn out to be not very different: compare Figures 1 and 6 of Zhang & Cottrell (2005). Interestingly, Freiwald et al. (2009) found that the macaque middle face patch is sensitive to configural changes, even though it is fairly early along the visual processing pathway. This suggests that it might be involved in both detection and identification. Therefore, it seems likely that it is the response properties of face representation units that is paramount for understanding holism, while the usage of these units for detection or identification may be secondary.

3.6 Spatial frequency

There is no agreement among models of face processing regarding the relationship between holism and spatial frequency, similar to what we find from empirical studies (see Chapter 9). Several models do not consider the issue at all (e.g. Jiang et al. 2006, Rossion 2008), perhaps considering it irrelevant. Among the models that do, some favor low spatial frequencies (LSFs), others favor high spatial frequencies (HSFs), but some are also neutral or ambivalent.

Among the models favoring LSFs, only Dailey & Cottrell (1997, 1999) explicitly examined the issue. When two competing modules were fed information from different SF bands, the module that received LSF information showed a strong face specialization. Interestingly, this only occurred under specific task conditions (subordinate classification for faces and superordinate classification of objects). However, even when both modules received identical information, face specialization also developed, albeit less strongly. The qualitative model of Tsao and colleagues (Tsao & Livingstone 2008, Tsao et al. 2010) also somewhat favors LSF. Their "obligatory detection stage… uses a coarse upright template to detect wholes faces" (p.421 of Tsao & Livingstone 2008). However, it is unclear if "coarse" necessarily implies LSF. Furthermore, usage of the term "coarse" in their descriptions of the model is sporadic, suggesting that they may not consider it to be a key aspect.

HSFs were not explicitly endorsed by any model. However, among the patches found to be best for face identification (highest mutual information) in Zhang & Cottrell (2005), the top 6 patches were all of mid to high spatial frequency (see Figure 6 of Zhang & Cottrell 2005). Of these top 6 patches, 4 were from the second-highest SF (out of five SF bands). Nonetheless, a closer examination shows that LSF patches had fairly high mutual information also. The model of Biederman & Kalocsai (1997 and McKone & Yovel (2009) similarly do not explicitly favor HSFs, but they argue that processing for faces "must encompass detailed local shape information", and "holistic processing derives from elements more local than the major parts", suggesting that their model would favor the usage of fine HSF information.

Some models are neutral or ambivalent on the issue of spatial frequency. The model of Schwaninger and colleagues (Wallraven et al. 2005, Schwaninger et al. 2009) combined LSF ("configural") and HSF ("component") information, stating that the combination of both types of information is what produces holistic representations. The patches of faces found to have the highest mutual information by Ullman and colleagues (Ullman et al. 2002) included a range of SFs (but note that the exact numbers were not stated; our interpretation is based on visual judgment from Figure 1 of their paper). However, among the top 8 patches, only one patch (ranked 5th) was unambiguously LSF. Interestingly, there seems to be some tradeoff between patch size and resolution (the LSF patch encompassed the whole face, while other patches were smaller), a tradeoff that we assume in our work. Zhang and Cottrell (2006) examined the effect of spatial frequency on their model's results, and found counter-intuitively that the ability to discriminate configural changes increased when higher SFs were used, because LSF templates are more shift-invariant. Nonetheless, their results matched human performance best at the 2nd-coarsest scale tested. The authors state that the issue requires "more careful treatment before [they] can draw any firm conclusions" (last paragraph, p.2433).

Overall, none of the models examined the issue of spatial frequency in detail. An important aspect may be the issue of cycles-per-face versus cycles-per-degree (or cycles-per-pixel, in the case of models). Many papers are not explicit about the exact numbers, therefore it may be difficult to make direct comparisons (LSF in one paper might only be MSF in another).

3.7 Configural versus featural processing

There is debate over whether inversion disrupts "configural processing" more than it disrupts "featural processing". Configural processing usually translates to sensitivity to distances between semantic face parts, while featural processing usually means sensitivity to swapping a face part from one person with that from another person.

According to the "perceptual field" hypothesis (Rossion 2009), the perceptual field (the "area of vision where… information for the task [can be extracted]", p.305) is constricted for inverted faces, and because of this reduced spatial window, each facial feature has to be processed sequentially and independently. Thus, configural processing is more disrupted than featural processing.

On the other hand, McKone & Yovel (2009) argue that the existing body of evidence indicates that featural processing is only less-disrupted than configural processing when "feature" changes included color and brightness (i.e. features that are not face-specific). When changes were only in terms of shape, both featural and configural processing are equally disrupted. Accordingly, McKone & Yovel propose that face-specific processing encompasses not only spacing between semantic parts, but also detailed shape information.

Similarly, Riesenhuber et al. (2004) also found that configural and featural processing are equally disrupted by inversion. Jiang et al. (2006) subsequently showed that a purely shaped-based model could be fitted to reproduce the results of Riesenhuber et al. (2004). Their results suggest that "configural" and "featural" may be an artificial distinction that may stem from the stimulus manipulations, rather than different underlying processing. This view mirrors that of Perrett & Oram (1993) and Farah et al. (1998); the latter proposed that holistic representations implicitly contain both first-order and configural features (p.495).

Zhang & Cottrell (2004) found that their Gabor-PCA model was overly holistic and did not match the human developmental data. By introducing a part representation, this problem was fixed. However, Zhang & Cottrell (2006) later found that the developmental data could be explained as a function of number of training images, or also as a function of increasing HSF information. Together, these findings suggest that their modeling results must be interpreted cautiously: there may be more than one model that can account for the data, and simply replicating the data is insufficient to distinguish between correct and incorrect ones.

Unlike the above models, the model of Schwaninger & colleagues (Wallraven et al. 2005, Schwaninger et al. 2009) was not tested on inverted faces, but on scrambled and blurred faces. In

their model, featural processing is related to high spatial frequencies, while configural processing is related to low spatial frequencies (this is the opposite of Zhang & Cottrell 2006).

Overall, there are three concrete models that claim to replicate the data for configural versus featural processing (Jiang et al. 2006, Zhang & Cottrell 2006, Schwaninger et al. 2009). However, there is disagreement as to whether these forms of processing are even different to begin with – and if they are, what the source of the difference is.

3.8 Face space and norm-coding

The notion of face space is basically the idea that individual faces can be represented as points in some multi-dimensional space. In theory, objects can also be represented as such. However, because faces share a common arrangement of semantic parts ("first-order configuration"), the notion of a face space appeals to the idea that perhaps the dimensions are "meaningful" ones such as second-order metrics (e.g. eye separation, nose-mouth distance) or first-order metrics (e.g. nose width, eyebrow angle).

Face space was first proposed by Valentine (1991). Specifically, two models were proposed: a norm-based coding model and an exemplar-based coding model. The norm-based model accords special status to the norm (average face), as representations are calculated with respect to the norm. In the exemplar-based model, representations are based on distance from some set of exemplars. Interestingly, both models accounted for the distinctiveness effect (better recognition for distinctive than typical faces) the same way – by relying on the assumption of a normal distribution. It is only for the opposite effect of distinctiveness for face detection that the two models have differing explanations. Ultimately, however, the two models were unable to be distinguished empirically.

Models of face processing differ with regard to norm-coding mainly in two aspects. The first is whether there is an explicit norm or an implicit one. The other is whether deviation from the norm is coded by opponent-pairs, where each half of a pair of neurons (or population of neurons) is directly opposite from the other (with respect to the norm). We discuss these two aspects next.

By default, most models that do not address the issue of norm-coding (e.g. Biederman & Kalocsai 1997, McKone & Yovel 2009, etc…) can be thought of as favoring an implicit norm. One idea, as articulated by Jiang et al. (2006), is that "there are more face neurons tuned to 'typical' faces that are close to the mean than to less common faces that are farther away from the mean" (p.168), mirror the assumption of normal distribution by Valentine (1991). Lewis & Johnston (1999) similarly utilized this assumption to account for the caricature effect without an explicit norm (see next section).

Rhodes & Jeffery (2006) proposed that the norm is implicitly represented, and identity is coded by opponent neuronal populations. For each dimension of face space, there is a population of neurons that responds strongly to low values for that dimension, and there is an opponent population that responds strongly to high values. Thus, the norm is implicitly coded, by equal activation of both populations.

Another method of coding for a norm implicitly is via competitive, "repulsive" interactions between representational units that cause the units to "move away" from each other in face space, so that they are equally spaced out. One such model is that of Brunelli and Poggio (1993), which is discussed in the context of caricatures in the next section. This method is compatible with the idea of opponent populations, although it may not necessarily produce units with exactly opposite tuning properties.

The influential "eigenfaces" method (Turk & Pentland 1991) is one that uses an explicit norm. During the PCA process, the mean face is explicitly subtracted, and the directions of largest variance around the mean face are calculated. While studies have tried to determine if the principal components actually correspond to the dimensions that human face perception is attuned to (e.g. Hancock, Burton & Bruce 1996), it seems that the results may not be conclusive, because the principal components depend heavily on the set of faces used.

Overall, apart from the caricature effect (next section), there do not exist any implemented models that account for face-space and norm-coding data, such as adaptation aftereffects (e.g. Leopold et al. 2001, Rhodes et al. 2004) and neural response properties (e.g. Leopold et al. 2006, Freiwald et al. 2009).

3.9 Caricatures

The notion of caricatures stems from face space and norm-coding. During the process of creating a caricature, the features of a given face are exaggerated away from some typical or average value (e.g. nose enlarged, eye separation increased). Thus, caricaturization can be thought of as shifting a given face away from the norm, especially for the dimensions in which the face is already atypical. For example, a face with big eyes is given even bigger eyes.

Brunelli & Poggio (1993) represented faces by a vector of automatically-extracted features such as nose width, mouth height and lip thickness. To perform gender classification, they created a "HyperBF" network in which two competing prototypes (each representing the average vector for one gender) were modified during a learning process. This process resulted in altered prototypes that enhanced the differences between the genders. In other words, the prototypes started off as the average male and female faces, but became caricatured male and female faces. Crucially, there was no explicit representation of the overall average face (norm), yet the prototypes seemed to move away from this implicit norm in directions that specifically enhanced their male-ness and female-ness. This scheme was then extended to the identification of individuals, with one prototype per individual. Again, the learning process resulted in caricatured prototypes. However, identification performance was already at ceiling even when the original, non-caricatured prototypes were used. Therefore, it is unclear if having caricatured prototypes is actually beneficial for identification of regular faces. Nonetheless, it is clear that caricatures would be better recognized than regular non-caricatured faces, since caricatures would activate the prototypes more strongly.

Lewis & Johnston (1999) combined the assumption of Valentine (1991) of a gaussian face-space distribution with an exemplar-based model featuring "Voronoi" cells. Each cell contains the space around one exemplar, and points in space that are equidistant from different exemplars form the cell boundaries. Because the gaussian distribution implies a higher exemplar density near the center of face space (i.e. the implicit norm), the exemplars are typically off-center within the cell, and are closer to the norm than the cell centroids are. Then, faces near the cells centroids are further away from the cell boundaries and therefore less prone to "jump" to an incorrect cell due to noise perturbations, hence these are better recognized. Since such faces are caricatures by definition (i.e. further away from the norm than the regular exemplars), the caricature effect is thus explained without an explicit norm.

Costen et al. (1996) show that by separating out the shape and texture components of a training set of faces and then performing PCA on these separate components, the resulting representation replicates the caricature effect. Unlike Brunelli & Poggio (1993), it is unclear if the prototypes (i.e. principal components) themselves are caricatured. More importantly, PCA utilizes knowledge about the average face, so the norm is explicit in this case.

Giese & Leopold (2005) tested two models: one prototype-based (like Brunelli & Poggio) and one that explicitly calculates distances from a face norm. Both models were equally able to replicate the electrophysiological findings for "lateral" caricatures (Leopold et al. 2006), but the norm-based model was better for "normal" caricatures. Thus, the results suggest that the norm-based model is a better representation of the actual neurophysiology underlying caricatures.

Overall, there exist multiple models that implement or replicate data relating to caricatures. This is somewhat surprising, given that only a handful of implemented models are geared towards replicating the FIE and CFE. In any case, however, the most important point is that these models are distinct – there is no model that can explain or link all of these phenomena.

3.10 Contrast polarity

Few models, either qualitative or quantitative attempt to account for the sensitivity of face processing to contrast polarity. This is surprising, since this is a well-established phenomenon specific to faces, but not objects (see Chapter 8). Two such models are described here.

The model of Biederman & Kalocsai (1997) utilizes Gabor-jets as the basic representation for faces, while also representing the spatial relationships between these jets. Sensitivity to contrast polarity is not actually demonstrated, although they hypothesize that the Gabor-jet representation makes the model sensitive to changes in lighting direction and contrast reversal. Crucially, the difference between the contrast sensitivities of faces and objects is attributed to objects being represented as a "structural description specifying qualitative characterizations" based on edges and depth discontinuities. These characterizations have been abstracted away from their initial Gabor-jet representations, and are therefore insensitive to contrast reversal. The reason for this difference in representation is attributed to faces requiring "highly accurate representations" for "accurate storage of and discrimination among thousands of highly similar exemplars" (Biederman & Kalocsai 1997, p.59).

Sinha (2002) performed face detection via multiple binary contrast polarity relationships between pairs of facial regions. In face images, the eyes are generally darker than the cheeks, and the mouth is generally darker than the chin. The proposed model searches for such binary contrast relationships in an image, and the presence of several of these suggests that the image contains a face. Because the representation is binary-valued (i.e. "darker-than" or "lighter-than"), this method is tolerant to changes in lighting direction and overall illumination level, while being sensitive to contrast polarity. However, this method is proposed as a generic recognition scheme (i.e. not face-specific), and face detection was simply the task chosen for demonstration purposes. Thus, it is not clear why faces and objects would differ in their sensitivity to contrast polarity under this scheme. One possibility is that this model only works for objects with a relatively uniform first-order 3D structure or uniform pigmentation scheme.

Overall, there are few satisfactory accounts of the sensitivity of faces (and insensitivity of objects) to contrast polarity. In contrast (pun not intended) to the two models discussed here, our account for this phenomena is based on the coarse spatial scale of face templates (see Chapter 8).

3.11 Neurophysiology

To date, there have been many models that attempt to account for behavioral phenomena associated with face processing. However, few – perhaps even none – have really tried to account for face processing at the neural level. The main reason is probably that until recently, face-selective cells could not reliably be found, and therefore many electrophysiology studies examined only a limited number of such cells. Thus, it was not always clear if the data were robust enough to try modeling concretely. However, with the recent advent of fMRI-targeted electrophysiology, robust data is no longer an issue. In this section, we briefly examine some electrophysiological findings and see if the above-mentioned models are consistent with them.

The neural FIE – a reduction or delay in response when faces are inverted – is still not well characterized. Early studies found different results in different brain areas. A recent study (Tsao et al. 2006) found that overall, a strong 2.2x (55%) reduction in response. However, only one face patch was studied. Furthermore, the distribution of response reduction was not examined, so it is unclear if the reduction is similar for all neurons. All models that account for the behavioral FIE can account for the reduction in response quite straightforwardly. However, a delay in response was also found, and models generally do not have good accounts of the dynamics of face processing.

Freiwald et al. (2009) found that in the middle face patch (sub-patches "MF" and "ML") of macaques (Moeller et al. 2008), face-selective cells are tuned to the presence/absence of at most four (out of seven possible) face parts. No model has directly attempted to explain this, especially in relation to holistic face processing. However, the model of Jiang et al. (2006) and Rossion and Gauthier (2002) are potentially consistent with these results. It is unclear if the Gabor-PCA models (which perform PCA on the entire Gabor response pattern to a whole face) are consistent or not – it may vary with the specific principal components. Generally, since most

models do not specify which brain areas correspond to which parts/aspects of the model, it is hard to really test or falsify these models.

Freiwald and colleagues (Freiwald et al. 2009) also found that face-selective cells in the middle face patch are tuned to the spacing of semantic face parts. Moreover, cells were most often tuned to the extremal spacing values. This is most directly related to the opponent-coding model of Rhodes & Jeffery (2006), which proposed two opponent populations of cells for each dimension of face space. However, it is unclear if other models are consistent or not with these findings.

Finally, the study of Kobatake & Tanaka (1994) had a cell that seemed to exemplify holistic behavior (Tsao & Livingstone 2008, Tsao et al. 2010). This cell responded strongly to a cartoon face (Fig 3.1 B), but not when the internal features were missing (C, D, F). The cells was only somewhat sensitive to the background luminance (G), but was highly sensitive to the luminance in the eye and mouth regions (H). Crucially, the cell required a face boundary to respond strongly (E vs. G), and this was taken as evidence for "Gestalt" or "holistic" processing (Tsao et al. 2010). While this was just one cell – and it seems to contradict the finding that middle face patch cells are sensitive to four face parts or less (Freiwald et al. 2009) – it is interesting to see if any model can replicate this behavior. The responses of the cell seem to be highly non-linear, so it is not clear if linear models (such as the Gabor-PCA models) can account for them. The non-linear tuning of the model in Jiang et al. (2006) may be more suitable, but this has not been tested.

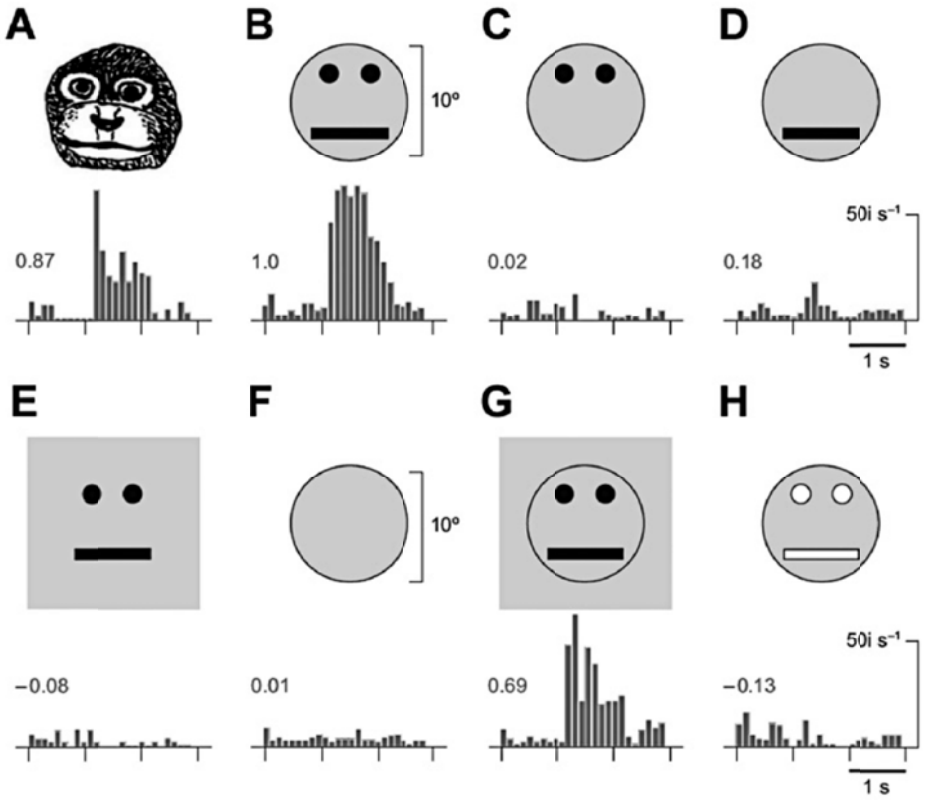


Figure 3.1. A "holistic" face cell found by Kobatake & Tanaka (1994). Numbers at left of PSTHs indicate normalized firing rates. Figure from Tsao et al. (2010).

3.12 Gabor-PCA model

Finally, we devote a section to the Gabor-PCA model of Cottrell and colleagues (Dailey & Cottrell 1999, Cottrell et al. 2002, Zhang & Cottrell 2004, 2005, 2006). Among the models reviewed here, the Gabor-PCA model has been the most extensively applied to various aspects of face processing. It is also somewhat similar to the model used in this thesis. Two key differences are in the template-matching process and the use of spatial frequencies. Here, we simply aim to put together in one place, a review of the findings from this model, as these have been scattered throughout this chapter.

The replication of the CFE is the most relevant finding. This was done for both "partial" (Cottrell et al. 2002) and "complete" (Richler et al. 2007) designs. However, the CFE was not investigated for inverted, SF-filtered, contrast-reversed, or non-faces. Moreover, it is not clear exactly how holism arises – whether it is due to the PCA, to the usage of all SFs, or the usage or all locations – and when it doesn't.

The relationship between holism and face identification is investigated in Zhang & Cottrell (2005). They found that the optimal face patches for face identification were not whole face patches (which were used in Cottrell et al. 2002), but covered at least two semantic face parts. They interpret their results to suggest that holism arises because it is good for face identification. However, they did not actually show that these optimal face patches were in fact holistic i.e. able to reproduce the CFE (in fact, Cottrell et al. 2002 was not even cited, for some reason). Interestingly, the optimal patches were all from high to mid SFs (in contrast to our work).

On a related note, Dailey & Cottrell (1997, 1999) found that a specialized face-processing area arose only for the combination of a low SF bias and differential tasks demands (identification for faces; categorization for objects). This is somewhat contradictory to the results of Zhang & Cottrell (2005), in which the purportedly holistic face patches were biased to high SFs.

Finally, Zhang & Cottrell (2004, 2006) found that the Gabor-PCA model was able to replicate the developmental trajectories of configural versus featural processing, and the effect of inversion on these. Interestingly, their results suggest that the increased discriminability of configural changes over time, is due to the increasing availability of higher SFs. While they explain that LSFs are more shift-invariant and therefore less suitable for detecting configural changes, it is still not clear why HSFs would favor configural over featural processing.

To summarize, while the Gabor-PCA model has been applied broadly to several interesting aspects of face processing, there remain two major issues. Some of their results seem to be somewhat contradictory, and have not yet been reconciled (to the best of our knowledge). Secondly, these results have used slightly different variants of the model, and it is not clear if a single model would be able to replicate all of their results.

3.13 Chapter summary

Overall, the state of models of holism (and of face processing in general) is rather unsatisfactory. Few have good accounts for the processing of inverted faces and non-faces, while none have really isolated what causes holism, e.g. by proposing a thorough mechanistic explanation for the CFE, or by demonstrating holism from the single cell level to the behavioral/perceptual level. More broadly, no single model ties together all the major aspects of face processing, such as holistic/configural processing, norm-coding, responses to inversion, SF filtering and contrast reversal, etc.

This thesis will attempt to address all the issues listed here. We begin by proposing in the next few chapters a thorough, step-by-step explanation of how the CFE arises.

Chapter 4: The HMAX model

Chapter abstract

The main objective of this chapter is to provide a brief description of the HMAX model of object processing, since this thesis builds on the HMAX model. Additionally, detailed methods for the experimental simulations in the various chapters are included here.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 4: The HMAX model

The objective of this chapter is to describe the overall operation of the HMAX model of visual object processing, which is the key simulation tool used in this work. HMAX models the large-scale neuronal operation of the ventral visual cortex in primates. It is a "neural-network" model in some senses of the term, but unlike other models such those of Dailey & Cottrell (1999) or Rumelhart & McClelland (1986), its origins are in Neuroscience rather than Cognitive Psychology.

4.1 Brief history

HMAX (Riesenhuber & Poggio 1999) was designed to demonstrate that a quantitative model could replicate several key characteristics of neurons in ventral cortex. Such neurons form a hierarchy, and neurons in "higher" areas are tuned to more complex visual stimuli (Felleman & Van Essen 1991). At the same time, neurons in higher areas are more tolerant to changes in position and size of visual stimuli (Tanaka 1996, Hung et al. 2005, Rust & DiCarlo 2010). Finally, neurons near the top of the hierarchy are view-tuned, rather than view-invariant (Logothetis & Pauls 1995, Logothetis et al. 1995).

HMAX is not completely unique; it shares many characteristics with other models of visual processing (e.g. Fukushima 1980, Perrett & Oram 1993, Mel 1997, Wallis & Rolls 1997, Amit & Mascaro 2003, Ranzato et al. 2007). As such, we often refer to these models as a "family of models" that take inspiration from primate visual neurophysiology.

The parameters were initially manually determined. Serre & Riesenhuber (2004) later modified these parameters to be quantitatively consistent with the known values in published neurophysiological studies, wherever possible. Nonetheless, there are still many free parameters unconstrained by actual data.

4.2 Implementation details

There are three aspects to the implementation of HMAX: general architecture, neuronal operations, and numerical parameters. The general architecture has already been described above: a hierarchical structure, with units of increasing selectivity for complex stimuli and tolerance to position and scale. More specifically, the hierarchy consists of layers that either increase selectivity or tolerance, in an alternating fashion. These layers achieve these different goals by performing different neuronal operations, described in the next section. Finally, both the general architecture and the neuronal operations have parameters that need to be specified.

In this work, the software implementation of HMAX that we used is the "HMAX package" within the Cortical Network Simulator (CNS) software system developed by Mutch et al. (2010). CNS is a broad simulation framework that is capable of simulating many different architectures. We used the "PNAS parameter set" within the "HMAX package" of CNS, meaning that general

architecture and neuronal operations follow that of Serre et al. (2007). The numerical parameters also follow those of Serre et al. 2007 (i.e.) to the extent possible.

Importantly, this work uses the "PNAS parameter set" as-is, without any modifications. The sole caveat is that we adjust the template parameters in accordance with our hypothesis about what differentiates face and object processing (large/coarse versus small/fine templates). No other changes are made to make the HMAX computations face-specific (e.g. calculation of eye separation). The significance of this is that according to our work, the key difference between face and object processing may be a quantitative rather than a qualitative one.

We now describe the different layers in our hierarchical model.

4.2.1 Multi-scale image pyramid

The model takes as input images that are 256-by-256 pixels. This image is then resized into 10 different scales to form an "image pyramid" (See Fig ??.??). This multi-scale representation is common to many computer vision algorithms.

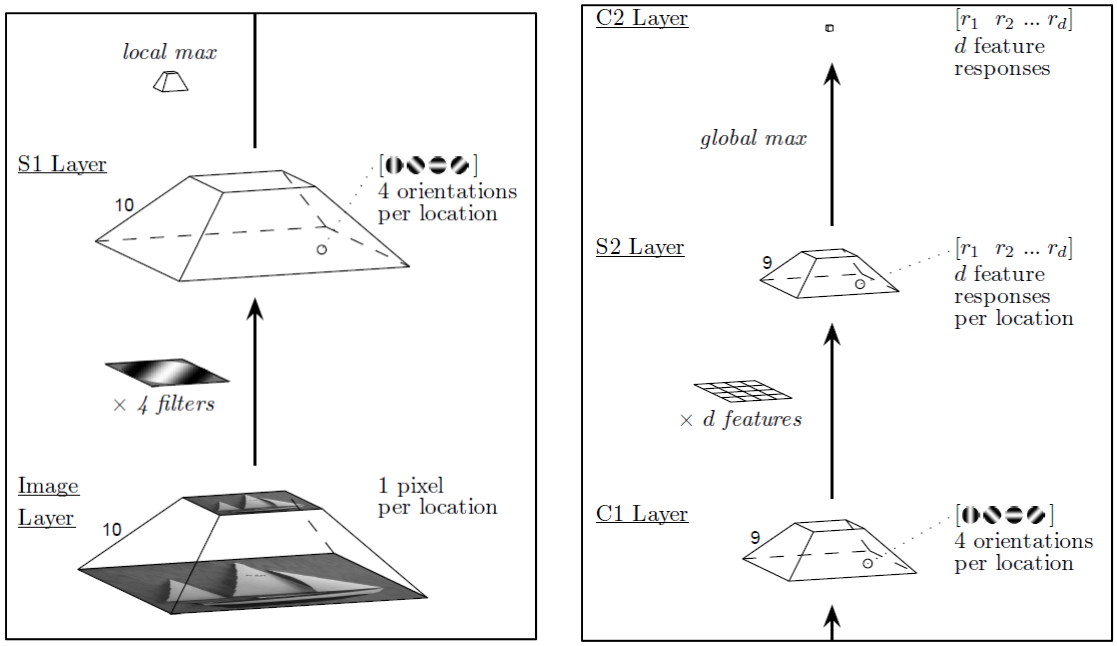


Figure 4.1. Schematic of the HMAX model. Processing sequence starts from bottom left ("Image Layer") and ends at top right ("C2 Layer"). Figure split into two to reduce space usage. Adapted from Mutch & Lowe (2006). © 2006 IEEE.

4.2.2 V1-like processing

Next, the model simulates the processing by simple and complex cells of the primary visual cortex ("V1") in primates. In the **S1** layer, template matching is performed on the image

pyramid. The matching uses the *normalized dot product (ndp)* function as a measure of how well each portion of the image pyramid matches the templates. Since the templates are Gabor filters at 4 different orientations, the template matching process is analogous to transforming the image into 4 channels, each of which corresponds to a different orientation. Therefore, each point in the pyramid now consists of 4 **S1** units. The response of each **S1** unit represents the amount of orientation information for a particular orientation, at a particular scale, at a particular location. This is similar to what simple cells in V1 do.

Next, in the **C1** layer, tolerance to position and scale changes in the image is increased by pooling over a small local region in the image pyramid. Specifically, every local region consisting of 8 x 8 **S1** units x 2 scales is pooled into a single **C1** unit by taking the maximum value over this 8 x 8 x 2 region. This is performed independently for each orientation. Since adjacent **C1** units have very similar values, for efficiency purposes, **C1** units at every other location are discarded ("downsampling"), reducing the number of **C1** units by a factor of 4. Note that compared to the **S1** pyramid, the **C1** pyramid is much smaller (as a combined result of pooling and downsampling).

4.2.3 High-level templates

Next, the **S2** layer again performs template matching (like **S1**). This time, instead of 4 Gabor filters, the templates can be an arbitrary number of arbitrary templates. In practice (e.g. Mutch & Lowe 2008, Serre et al. 2007), the templates are random "snapshots" of local regions (e.g. 4 orientations x 12 x 12) of **C1** unit responses at some scale. These **C1** responses are usually those elicited in response to the presentation of an arbitrary subset of the image set used.

In our case, out of the 100 frontal-view, oval-cropped male faces from the Max Planck dataset, we use the 50 odd-numbered ones (e.g. Fig. 4.2 top left) for this random template "snapshot" process. The other 50 are used to construct composites (e.g. Fig. 4.2 bottom left). For each of the 50 odd-numbered faces, 20 template "snapshots" are taken are random locations in the image, resulting in 1000 templates total.

For our face-like processing, these 1000 templates are of size 12 x 12 and are from scale 7 (where scale 9 is the coarsest). We call these "large, coarse templates". (The 4 orientations are implicit, i.e. the templates are actually 4 x 12 x 12)

For our object-like processing, the 1000 templates are also of size 12 x 12, but are from scale 3 (where scale 1 is the finest). We call these "small, fine templates".

It is important to note that the same 50 face images are used in both cases, even in the case of "object-like" processing. This is so that we can avoid the confound of different physical stimulus properties, which might arise if non-face images were used.

Where useful, we also utilize templates of size 24 x 24 from scale 3, and templates of size 4 x 4 from scale 7. We terms these "large, fine templates" and "small, coarse templates" respectively.

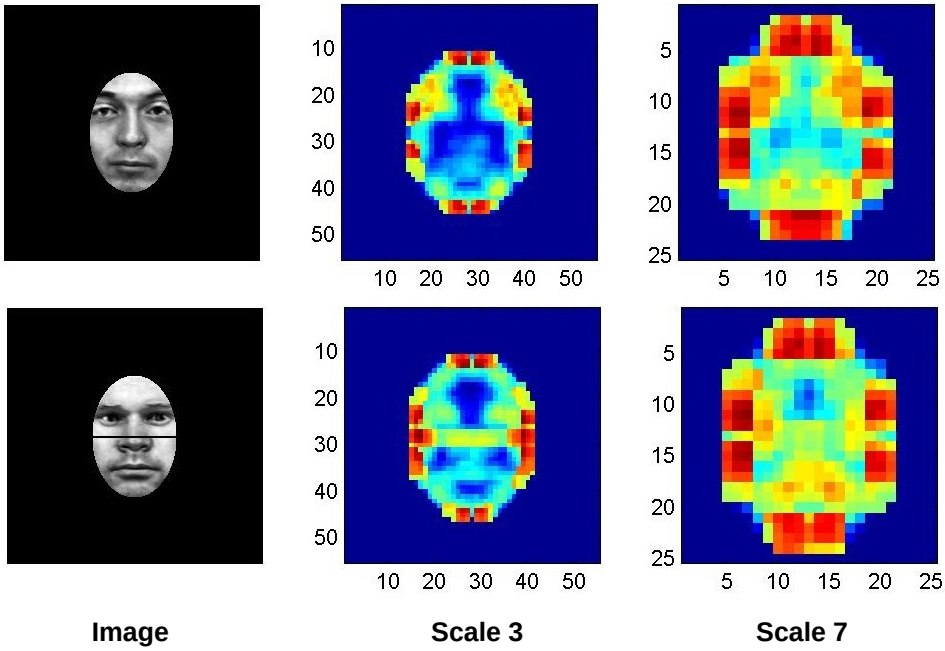


Figure 4.2. Left: examples of original (top) and composite (bottom) images. Middle: **C1** responses at scale 3. Right: **C1** responses at scale 7. Depicted response is the mean over 4 orientations. Blue: low response. Red: high response.

Note that "coarse" or "fine" refers to scales 7 and 3 respectively. "Large" and "small", however, do not simply refer to the numerical size of the templates, since the large, coarse templates and small, fine templates are all 12 x 12. Rather, they refer to what proportion of the whole face the templates cover. At scale 7, the entire image is represented as 25 by 25 **C1** units, and the whole face by approximately 22 by 17 **C1** units (height x width). Therefore 12 x 12 is a relatively large proportion. In contrast, at scale 3, the entire image is 55 by 55 **C1** units, and the whole face is about 36 by 28. Therefore, 12 x 12 is a relatively smaller proportion.

4.2.4 High-level template matching

The template matching process in **S2** is essentially the same as in **S1**. Note, however, that in terms of practical implementation, both Mutch & Lowe (2008) and Serre et al. (2007) utilize a *Gaussian radial basis function (grbf)* in **S2** to measure similarity, instead of the *ndp* function used in **S1**. Empirically, both functions give similar levels of object recognition performance. Comparisons of the two functions are beyond the scope of this work. We simply follow the design decisions made by Mutch & Lowe (2008) and Serre et al. (2007).

Note that each point in the **S2** pyramid corresponds to 1000 features (one feature is synonymous with one template). This is like how each point in the **S1** pyramid corresponds to 4 orientations.

4.2.5 Global pooling over location and scale

Finally, at the **C2** stage, we also perform pooling by taking the maximum value over a given region (analogous to the pooling in **C1**). However, this time, the pooling is done over all locations and all scales (i.e. the entire pyramid) – but separately for each feature, like for **C1**. This provides much greater tolerance to changes in image position and scale. Because pooling is done over the whole pyramid, the pyramid is reduced to a single point. However, because the pooling is done separately for each feature, these remain distinct. Therefore, the end result is that the pixel image is ultimately represented as a vector of 1000 feature responses. Each response is between 0 and 1, and can be thought of as the output of a graded feature detector (where the feature being detected is specified by the corresponding **S2** template) that is tolerant to position and size variations.

4.3 Detailed methods

4.3.1 Choice of scale and template size

We chose scales 7 and 3 to be distinct enough, yet not at the scale extremes (9 and 1). We chose the template size of 12 x 12, so that at scale 7, the templates would unambiguously not cover the whole face. This was to underscore the point that holism is not necessarily about wholes. We chose to also use 12 x 12 templates at scale 3 so that the "complexity" (number of **C1** inputs) is constant for both types of features.

For the small, coarse features, we chose the template size of 4 x 4 because it roughly corresponds to the size of semantic face parts (e.g. eyes, mouth), and therefore cannot be said to be too small.

For the large, fine features, we chose the template size of 24 x 24 arbitrarily to be "on the safe side", i.e. so that they would be large enough to also show "holism". We were cautious not to make the templates too large, as computation time could start to be prohibitively long.

4.3.2 Randomization

To provide an estimate of uncertainty, the reported mean values are the mean over 100 randomized runs, in all cases. In each run, a different random pairing of faces is done, so that the set of composites used in each run is different. Error bars are standard error of the mean (SEM) unless otherwise reported.

In most runs, all 1000 features are used. In some instances (e.g. in Chapter 5), a random subset of features is used in each run, to verify the robustness to number of features. In this case, randomization of face pairing (previous paragraph) and feature subset occurs together in the same run.

Our randomization is not meant to simulate experimental noise or inter-subject variability, so our measures of uncertainty (error bars) cannot be compared to empirical ones. We generally side-step statistical analyses in this thesis, because we believe these to be rather meaningless in the

absence of realistic sources of noise and variability. Reporting p-values (which would be extremely significant in almost all cases) might confer undeserved legitimacy. An effect that is statistically significant in our simulations could very well be non-significant in empirical studies. Except where noted, it is clear from the very small error bars (in most cases hardly visible) that effects are "statistically significant" in the regular sense.

4.3.3 Distance metrics

We assume that human behavioral responses in same/different tasks are generated based on the following intuitive description: the neural responses to each stimulus are compared using some distance metric, and if the neural responses are similar enough (i.e. distance is small enough), then the response is "same". It is unclear if this is what actually happens, but we make this assumption in the absence of compelling evidence otherwise.

Unlike the Euclidean distance, the Pearson correlation and normalized dot product (*ndp*) are similarity metrics, and attain their maximum value when the inputs are identical. To convert these into distance metrics, we simply calculate $1 - x$, where x is the correlation or *ndp*. Thus, a distance of 0 is attained when the inputs are identical. Note: the output ranges of [-1 +1] become [2 0] (or [0 2], rather) as a result.

4.3.4 Thresholds

In each randomized run, 20 trials for each of 16 conditions (upright/inverted x same/different x aligned/misaligned x congruent/incongruent) are generated. To determine whether our model responds "same" or "different" on each trial, a threshold is required. If the distance between the two composites in a trial is below the threshold, our model responds "same".

Instead of pre-determining a threshold, or trying to find an optimal one, we calculate results (e.g. hit rate, D', accuracy) for a range of thresholds spanning the entire range of distances in each run. The range of distances varies slightly from run to run. For each run, we linearly divide the range of distances by 18, thus providing 18 thresholds. Therefore, we are able to analyze the full range of possible outcomes independently of how thresholds are (implicitly) determined in humans.

In some instances (e.g. in Chapter 9), when we compare different sets of experiments (e.g. high versus low spatial frequency filtered faces), we run the simulations twice. The first time, the simulations are run as per normal, with thresholds determined as described above. We then examine the range of thresholds for each experiment to determine a common range that is suitable for all sets of experiments being compared. A tradeoff between coverage (covering all distances in all sets of experiments) and resolution (having thresholds that are not too far apart) is performed subjectively to decide the set of common thresholds to be used. A second round of simulations is then performed, with this set of thresholds used in all runs, for all sets of experiments.

4.3.5 Stimuli

The bulk of our simulations (e.g. in Chapters 5 to 10) used faces from the Max Planck database. In particular, we used the set of 100 frontal-view male faces. These faces were oval-cropped to remove external features. Each face was then normalized so that the mean and standard deviation of pixel values in all faces was the same. The background pixel value is 0 (black) unless otherwise stated. When creating composites, the bottom halves were shifted downwards by 2 pixels to create a gap.

4.3.6 Re-centering

To match many empirical experiments (including our own pilot experiments), in which subjects fixated the eye region in the top halves, the composite faces were shifted downwards by 30 pixels, so that the eyes were in the center of the image. This was also done during the feature extraction process (Section 4.2.3) to ensure consistency. Incidentally, this may have parallels in human vision, e.g. slight favoring of the eye region over the mouth region.

4.3.7 Attentional modulation

In the current absence of a good understanding of attentional modulation in humans (and primates), we simply perform the simplest version of attentional modulation by multiplying the pixels values corresponding to the lower face halve by some fraction. This sidesteps the issue of determining which units should correspond to lower or upper halves, which would arise if modulation was done at some level(s) higher than pixels (e.g. **S1**, **C1**, etc.), which would be making further assumptions about how attention operates.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 5: Modeling the Composite Face Effect (CFE)

Chapter abstract

In this chapter, we verify our main hypothesis, that the Composite Face Effect (CFE) can be produced by using large, coarse templates in the HMAX model. When small, fine templates are used, no CFE is found. We also show that the CFE is qualitatively robust to various low-level factors. There are two crucial contributions. Firstly, we provide a mechanistic, step-by-step account of the CFE, linking holism at the single-unit level to the misalignment effect at the behavioral level. Secondly, we show that the key factor in producing holism is the largeness of the templates, in terms of "image coverage" rather than number of pixels or units.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 5: Modeling the Composite Face Effect (CFE)

In this chapter, we present our main result, which is that a computational model can replicate the Composite Face Effect (CFE). Our hypothesis is that holism, which the CFE reflects, arises as a result of neurons being tuned to large portions of faces (henceforth "large templates"). To test this hypothesis, we took the HMAX model of object recognition and modified it to have large templates. This single change is the only one necessary to produce holism in the model.

First, we show that holism at the level of a single neuron can also be shown by a model unit tuned to large templates. Then, we also show that such model units can replicate the CFE. We thus demonstrate that these units are holistic, both at the single-neuron "brain" level and the behavioral "mind" level. Crucially, we trace the mechanisms that enable the model to produce holistic behavior from holistic neural responses, providing a step-by-step account of how holism arises. We thus bridge the explanatory gap between brain and mind for holistic face processing.

5.1 Replicating holism at the single-neuron level

We first perform a quick validation of our hypothesis that holism arises due to large templates by showing that a model unit that is tuned to the entire image can replicate a single-neuron demonstration of holism (Fig. 5.1). Because the unit's template encompasses the entire image, the removal of any portion weakens responses substantially (Fig. 5.1 b, c, d). Interestingly, even though the face outline is only a few pixels thick, its removal also weakens responses substantially (e vs. f), reminiscent of "Gestalt" processing. Importantly, it is easy to see why a model unit tuned to only a portion of the image ("small template") would not be able to replicate these results. For example, if the template did not cover the mouth, then removal of the mouth (b, d) would have no effect at all, contrary to the neural data. More modeling details can be found in Section 4.3.

5.2 Replication of CFE (misalignment effect)

In order to simulate the experiments in which human subjects participated, we also did the following. To simulate the attentional modulation that happens when subjects try to ignore the bottom halves, we multiplied the pixel values in the bottom halves by a modulation factor (default value is 0.1). To simulate the process of deciding if the two top halves in each trial are the same or not, we first calculated the distance between the two sets of HMAX outputs (default metric is Euclidean distance). Then, if the distance is below a certain threshold, the two top halves are considered to be the same. Section 4.3 describes these methods (as well as stimuli and model parameters) in more detail, and also justifies each of these design decisions.

We found that HMAX with large, coarse face templates replicated the CFE (misalignment effect). Specifically, for the "partial design" (see Section 2.1), the hit-rate increased when the bottom halves of the stimuli were misaligned (Fig 5.2 left). As a control, we re-ran the model,

changing only the template size (but not coarseness). Consequently, the model failed to show an increase in hit-rate when the bottom halves were misaligned (Fig 5.2 right).

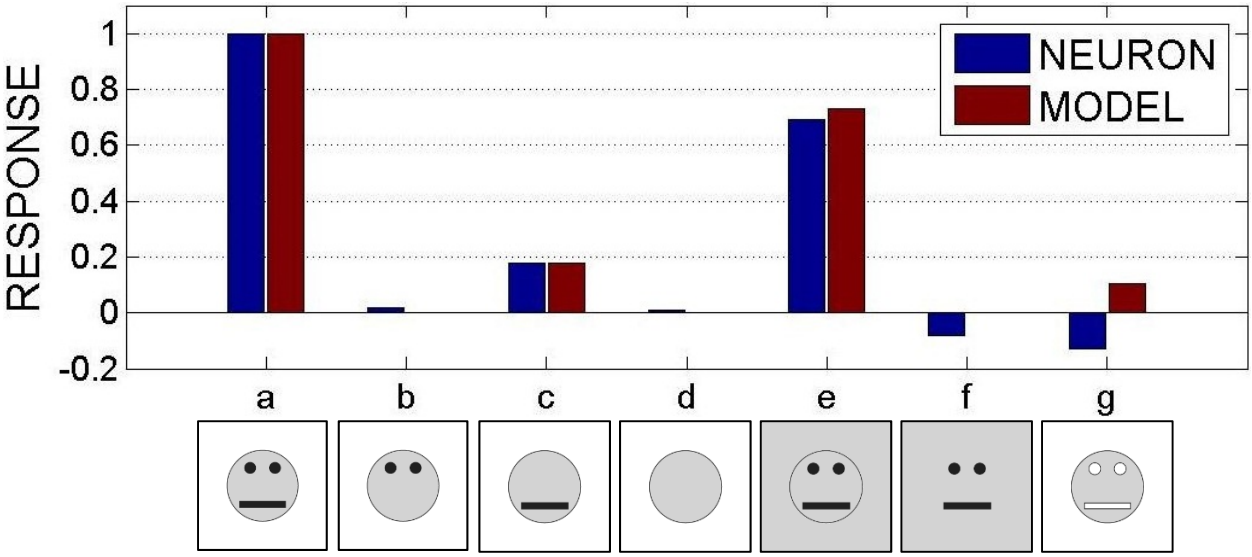


Figure 5.1. A model unit that is tuned to the entire image (a) is "holistic", responding significantly less when face parts are missing (b, c, d). The background contrast has only a moderate effect (e), but the face outline is essential (e vs. f), as is the right contrast polarity for the eyes and mouth (a vs. g). The model unit's responses are in good agreement with those of the neuron from Kobatake & Tanaka (1994, Fig. 4). Images reproduced from Tsao et al. (2010).

5.3 Step-by-step account

How does the model go from the holistic response properties of single units to produce a higher hit-rate for pairs of bottom-misaligned images? We first examine the responses of model units to misalignment of individual images, and then look at the subsequent effect on image pairs.

5.3.1 Effect of misalignment: individual images

Following from the holistic properties of units with large templates, we see that when the face halves are misaligned, these units respond less strongly (Fig. 5.3 left). In comparison, for units with small templates, the units respond as strongly as before (Fig. 5.3 right).

5.3.2 Effect of misalignment: distances between images

We now examine the effect of misalignment on the distances between the two images in each trial. Because misalignment causes responses to each image to decrease, this also generally causes distances between images to decrease (Fig. 5.4). In theory, this may not be true in all cases, but Fig. 5.4 shows that this holds true empirically. We speculate that the high

dimensionality of the response vectors (i.e. number of units), as well as the somewhat small and uniform decrease in responses (Fig. 5.3 top left) may contribute to this empirical result. However, further examination of the underlying cause(s) is beyond the scope of this thesis. The decrease in distance holds true even for a non-Euclidean metric such as the Pearson correlation (see Section 5.4.4).

The decrease in distances between images leads directly to an increase in hit-rate, if a relatively stable threshold in assumed (red line in Figs. 5.4 and 5.5). Since two images are declared to be "same" if their distance is below the threshold, a decrease in distance due to misalignment increases the proportion of image pairs below the threshold, i.e. a higher hit-rate. Note that the actual value of the threshold is not crucial, as long as it is fairly constant across trials.

Thus, we have traced the effect of misalignment from the level of a single model unit, to the "behavioral" level of producing a response of "same" or "different" (like human subjects). For large templates, there is a clear and step-by-step account of how holism at both levels is related. Correspondingly, for small templates, there is no evidence of holism (Figs. 5.2 right and 5.5).

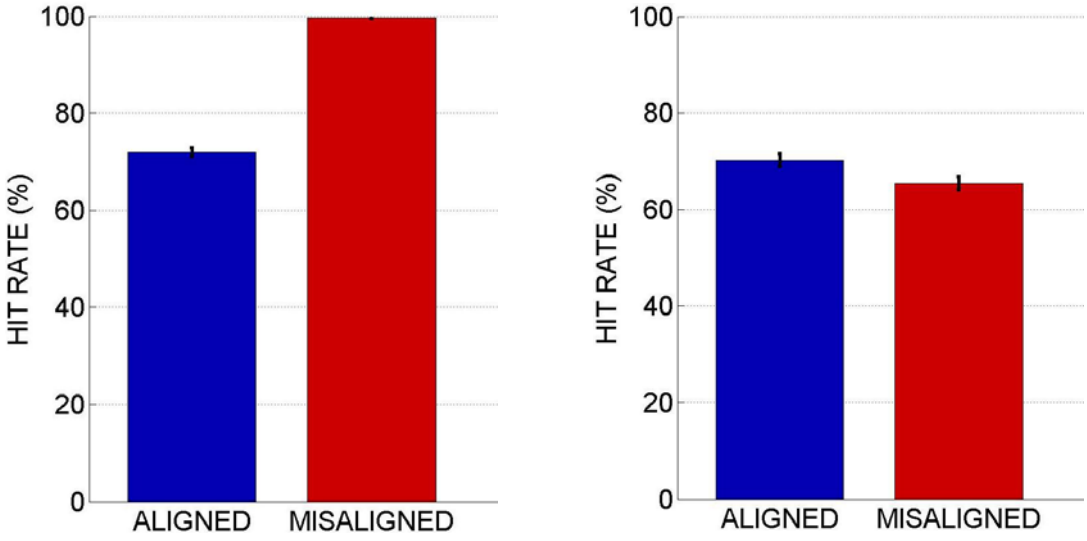


Figure 5.2. The model shows the CFE (misalignment effect) for large, coarse templates (left), but not for small, coarse templates (right). Error bars: ± 1 SEM.

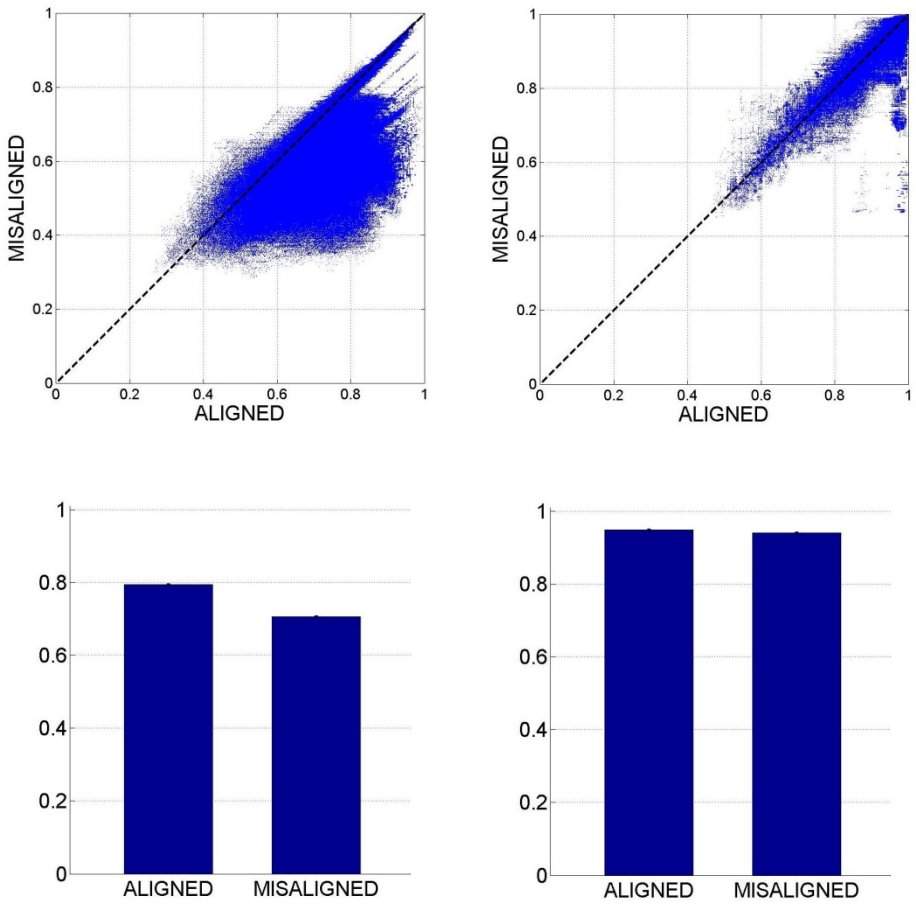


Figure 5.3. Misalignment causes decreases in responses for units with large (left), but not small (right) templates. Top: scatterplots for responses of 1000 model units to 2450 composites. Bottom: overall mean responses.

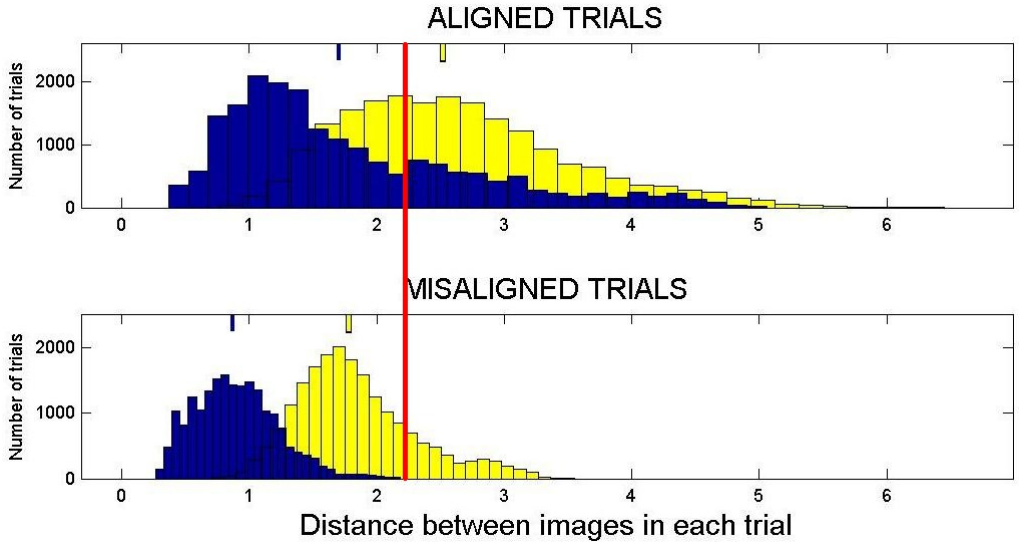


Figure 5.4. For large, coarse templates, distances between composites are larger for aligned (top) than misaligned (bottom) trials. Blue: "same" trials. Yellow: "different" trials. Hanging bar indicates mean of distribution. Red line: threshold producing roughly 75% hit rate for aligned "same" trials.

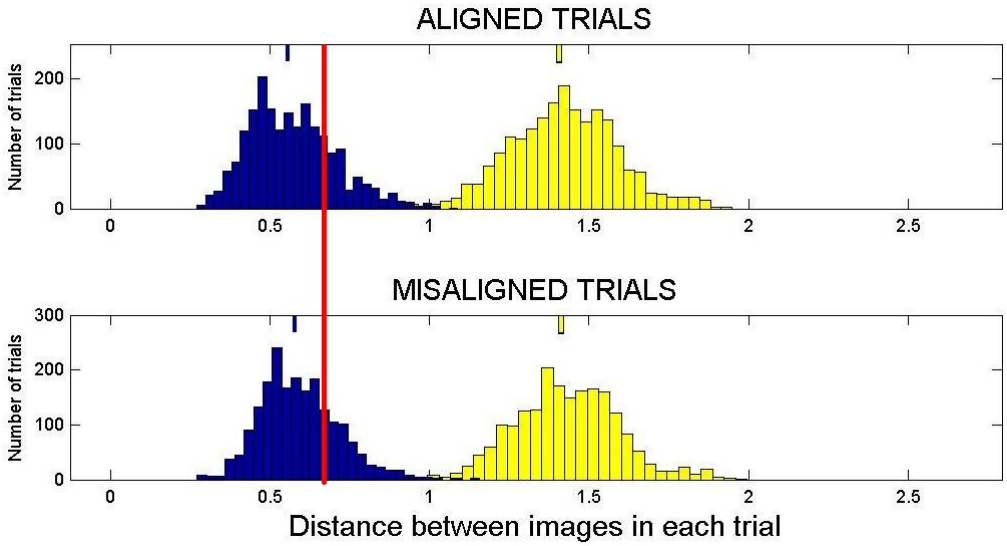


Figure 5.5. For small, coarse templates, distances between composites are similar for aligned (top) and misaligned (bottom) trials. Blue: "same" trials. Yellow: "different" trials. Hanging bar indicates mean of distribution. Red line: threshold producing roughly 75% hit rate for aligned "same" trials.

5.4 Robustness

Robustness to changes in parameter values is a desirable property for any model, qualitative or quantitative. Qualitative models sidestep this issue completely, but that does not mean that they are robust. On the other hand, quantitative models are forced to be explicit about design decisions and parameter values, allowing robustness to be evaluated. In this section, we show that our model is robust to various changes. These results can also be considered evidence that large templates are the key to holism, not other factors.

5.4.1 Threshold

The exact value of the threshold is not important for the model to show the CFE. Apart from extremely high and low threshold values leading to hit-rates of 0% or 100%, the model always shows an increase in hit-rate for misaligned trials (Fig 5.6). The amount of increase does vary, however, unsurprisingly. Note that for behavioral experiments performed to date, aligned hit-rates have ranged from 54% (Taubert & Alais 2009) to 88% (de Heering et al. 2007). However, since Gauthier and colleagues have shown that this can be manipulated (Richler et al. 2011b), the robustness of our results is reassuring. In this thesis, unless otherwise noted, results are shown for thresholds that produce aligned hit rates of roughly 70-85%, to avoid ceiling or floor effects.

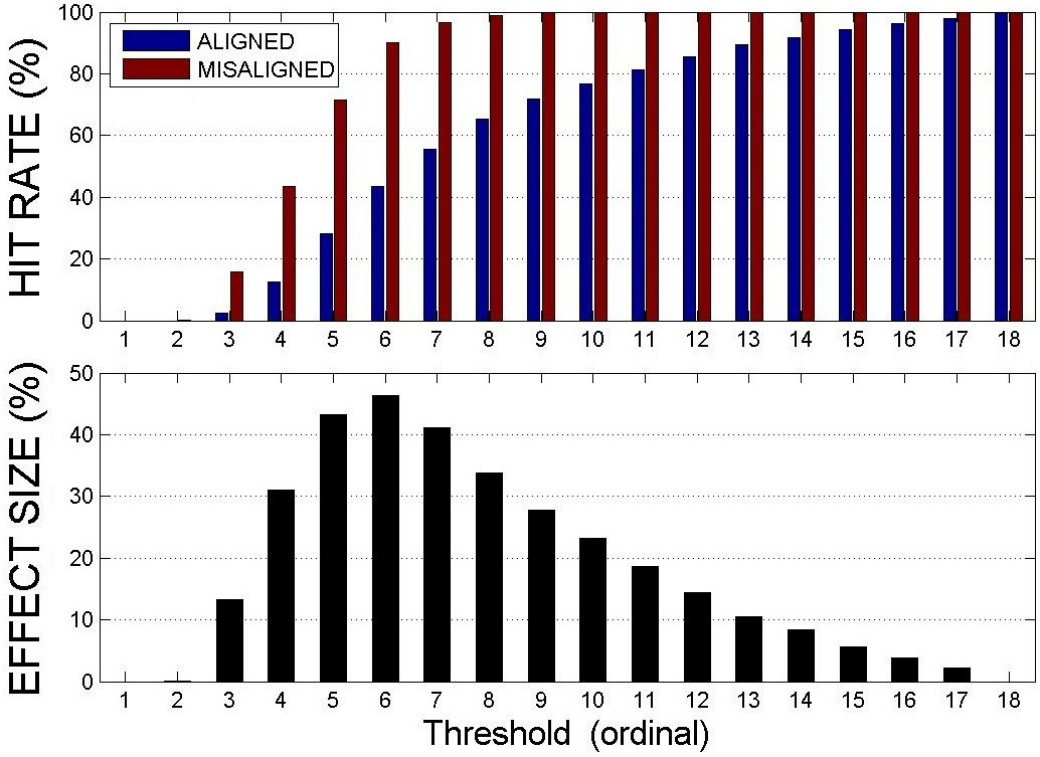


Figure 5.6. Top: hit rates for aligned (blue) and misaligned (red) trials as a function of ordinal threshold. Bottom: size of misalignment effect (misaligned hit rate – aligned hit rate) as a function of threshold. Note: ordinal, not numeric, threshold values are shown here, because exact numerical values vary slightly from one random run to another.

5.4.2 Jitter

In our pilot psychophysics (not reported here), we jittered the images to ensure that subjects did not perform the discrimination task by focusing on a small set of low-level features. This was not necessary for model simulation, since no low-level (i.e. V1-like) features were used. Nonetheless, the model is robust to image jitter (Fig. 5.7 left). This is not surprising, since the units are tolerant to position changes by design.

5.4.3 Number of features used

We have used 1000 model features as the default number. Figure 5.7 (middle and right) shows that the model replicates the CFE even with only a handful of randomly chosen features (out of the 1000) in each run, demonstrating robustness to the exact number of features used.

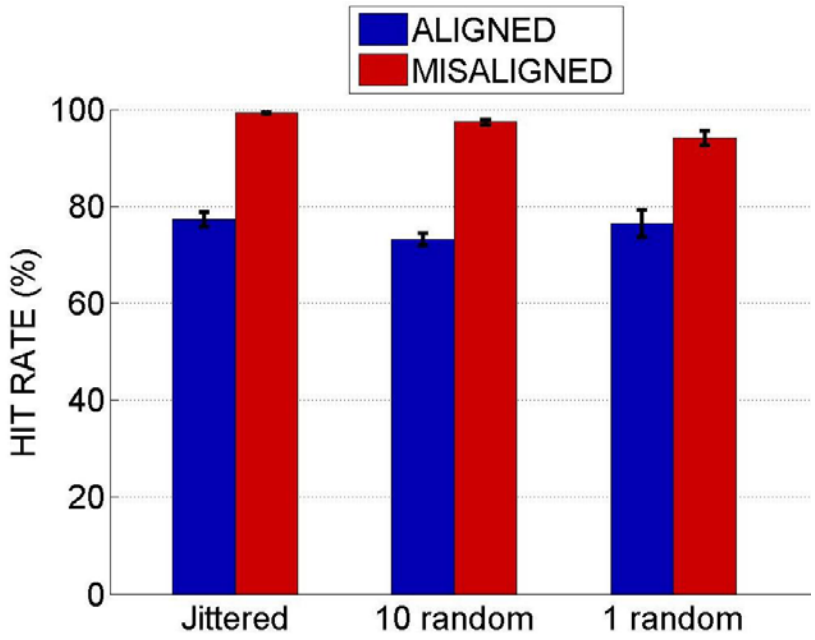


Figure 5.7. The CFE (misalignment effect) shown by the model (large, coarse templates) is robust to image jitter (left) and number of features (middle: 10 randomly chosen features per run; right: 1 randomly chosen feature per run).

5.4.4 Distance metric

Earlier, we noted that a decrease in response to an individual image does not necessarily mean that distances between images decreases. Nonetheless, this relationship is seen empirically, at least for the Euclidean distance metric. In Fig. 5.8 (left), we see that even for a non-Euclidean metric such as the Pearson correlation (r), this relationship holds. Thus, a misalignment effect (CFE) is also seen (Fig. 5.8 right)

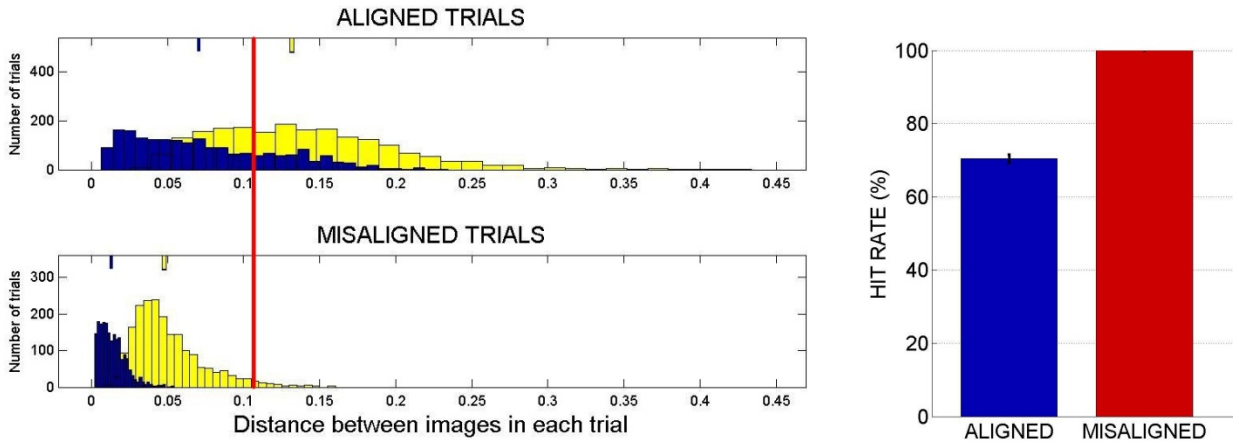


Figure 5.8. Using Pearson correlation as the distance metric (specifically, 1 – r), misalignment also reduces distances (left), creating a misalignment effect (right).

Since the Pearson correlation is invariant to linear transformations (e.g. subtractive or divisive decreases in this instance), one question might arise: shouldn't the decrease in responses (due to misalignment) be abolished by this invariance property? However, invariance is only true if all responses decrease by the same amount or same proportion. Since the decrease is not identical across features, thus there is no invariance, and the decrease in response affects the correlation.

5.4.5 Attentional modulation

We used multiplicative modulation of pixel values to simulate the attentional modulation achieved by subjects when told to ignore the bottom halves. The detailed mechanisms of attentional modulation are still unknown. Hence, we chose this method for its simplicity, despite of its non-realism. A similar method was used by Cottrell et al. (2002) and Richler et al. (2007).

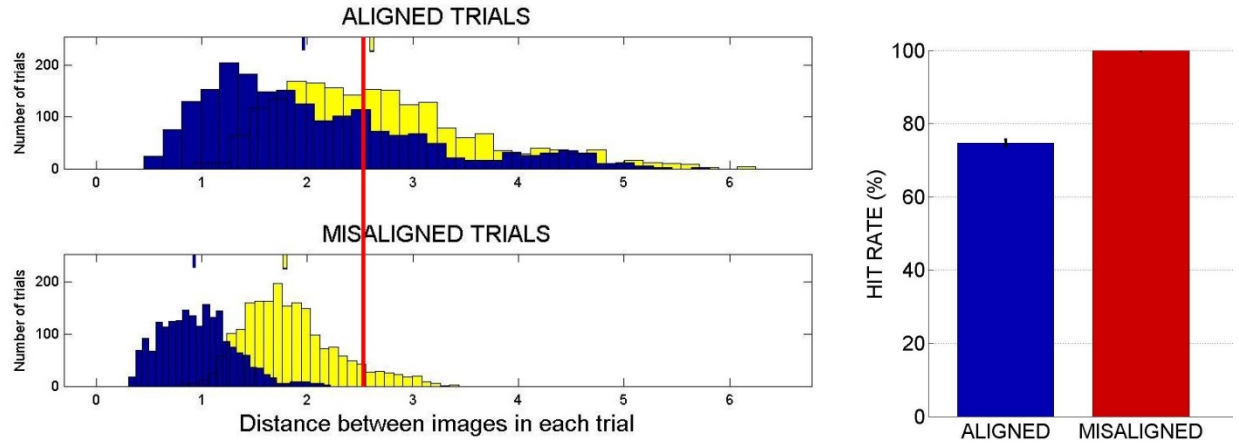


Figure 5.9. If pixel values for the bottom halves are multiplied by 0.5 (as used here) instead of 0.1, a misalignment effect is also seen, unsurprisingly.

Unless otherwise noted, we multiplied the pixel values of the bottom halves by 0.1. This is clearly more than the amount of modulation subjectively experienced by humans (empirically very moderate modulation, see Carrasco et al. 2000, 2004, Pestilli & Carrasco 2005). In other words, we have made it especially hard for our model to be affected by the bottom halves, i.e. to be holistic. Nonetheless, the exact amount of modulation is not crucial for the existence of the CFE (Fig. 5.9). Of course, in the extreme case (multiplication by 0.0), no holism is shown – nor should holism be expected.

5.4.6 Template matching

In our model, the S1 layer (approximating V1 simple cells) is derived from the multi-scale image pyramid by matching every possible region to Gabor templates. This template-matching process uses the *normalized dot product (ndp)* template-matching function to determine how well a given region matches the template.

There is also neurophysiological evidence that normalization actually occurs, at least in the cat (Heeger 1992) and primate (Carandini & Heeger 1994, Carandini et al. 1997) primary visual cortex. However, normalization counteracts the effect of the simulated attentional modulation (Fig. 5.10). Therefore, we need to discount the possibility that holism arises in our model only because the bottom halves are effectively restored to their full pixel intensities. (Although it should be noted, in any case, that small templates using the *ndp* do not show holism, Fig. 5.2 (right), so this control is not strictly necessary)

Using just the dot product (without normalization) as the function for template-matching, the model still replicates the CFE (Fig. 5.10 right). However, for a given level of attentional modulation (e.g. multiplication by 0.1), the effect is weaker when using the dot product (*dp*), as compared to using the *ndp*. Weaker modulation is required for a strong CFE to be shown.

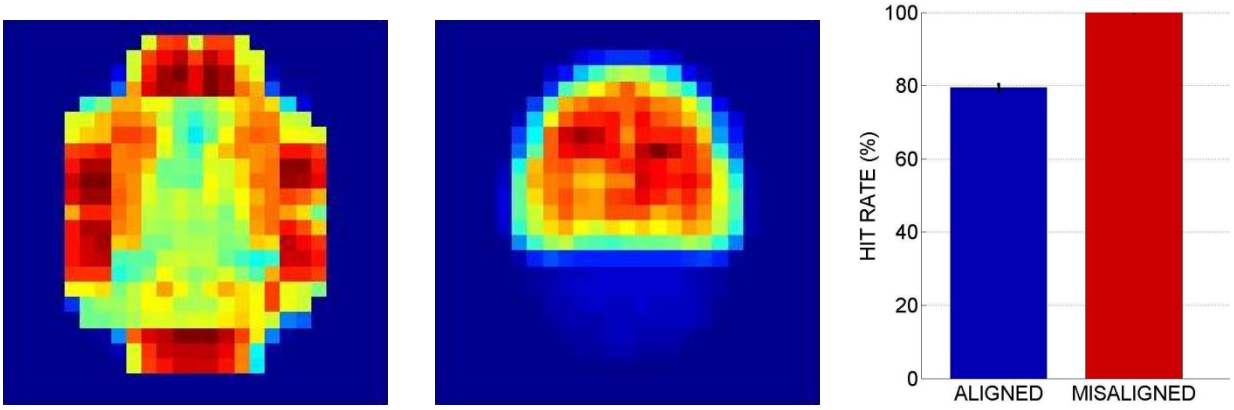


Figure 5.10. Using *ndp* (left), but not *dp* (middle) for template matching counteracts the effects of simulated attentional modulation. Mean C1 activity over all orientations is shown. (Note: color scales are different). Using a dot product, the model can still produce the CFE, but is less tolerant to attentional modulation. The CFE is strong for modulation of 0.5 (right), but less for 0.1 (not shown).

These results do not contradict our hypothesis or previous findings. As mentioned above, multiplication of 0.1 is a much stronger modulation than what humans experience. Furthermore, since normalization has been found in the brain, these results point to the unrealistic nature of our modulation method, rather than the problem with normalization. For the rest of this thesis, we retain the *ndp* as the template-matching function, in line with our philosophy of modifying HMAX only when absolutely necessary.

5.4.7 Background intensity

Unless otherwise noted, the pixel value of the background is set to 0 (black). As seen in Fig. 5.1, holistic face cells are slightly sensitive to the background value of an image. We wanted to verify that the background value during the feature extraction stage is not a crucial factor. This factor may seem unimportant currently, but it is important for the issue of contrast polarity (see Chapter 8). When the contrast of an image is reversed, it is an open question whether the background contrast should also be reversed or not.

When the background is set to 128 (halfway between black and white, i.e. grey), there are two effects. Firstly, the face boundaries are much less salient at all levels of representation (Fig. 5.11), unsurprisingly. Second, and more importantly, weaker attentional modulation is required for a strong CFE to occur (Fig. 5.12). This is reminiscent of the effect of using the *dp* instead of *ndp* for template matching (previous section). The underlying reasons are actually related. Since the background value is now much larger (128 vs. 0), the vector norm is fairly constant (and large), regardless of attentional modulation (which is applied only to the face). Hence, the amount of normalization (i.e. re-scaling to produce a norm of 1) does not depend on attentional modulation. In other words, normalization does not counteract attentional modulation, effectively reducing *ndp* to a *dp* (qualitatively, not quantitatively) when the background is grey.

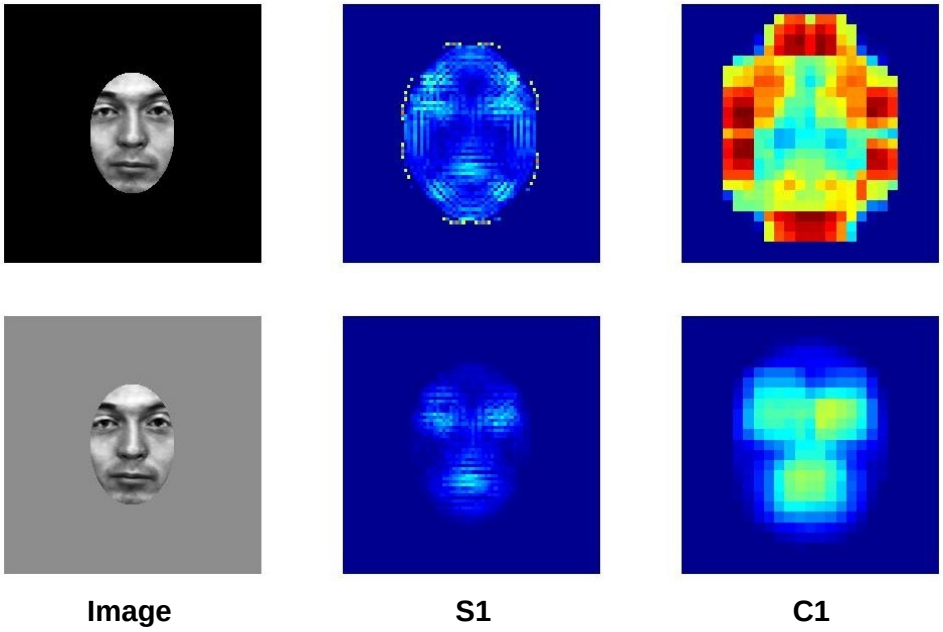


Figure 5.11. For a grey background (bottom row), the face boundary is perceptually less salient, compared to a black background (top row). This effect carries over to the S1 (middle column) and C1 (right column) activity, where the boundary is no longer apparent. Mean activity over all orientations is shown. Blue: low activity. Red: high activity.

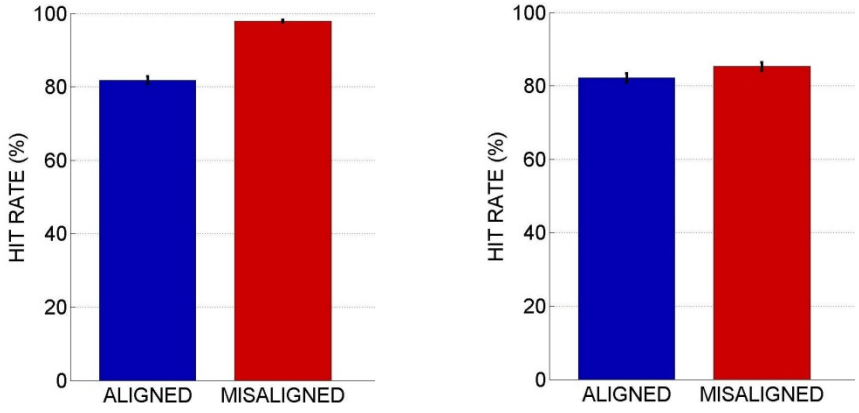


Figure 5.12. For grey backgrounds, attentional modulation of 0.5 produces a strong CFE (left), whereas modulation of 0.1 produces a very weak effect (right).

5.5 Factors affecting holism

Thus far, we have examined some low-level factors which the model is relatively robust to, in terms of qualitatively producing a misalignment effect. In this section, we examine factors that are more intuitively relevant to holism (this distinction is subjective, however). Recall that our hypothesis is that template size is key – large templates produce holism, small templates do not.

It is important to define template size more precisely. In the model, the spatial scales differ from one another in terms of their spatial frequency (SF) content. The finer scales represent information at higher spatial frequencies (HSFs); coarser scales represent lower spatial frequencies (LSFs). Because of certain implementation specifics, the finer scales have fewer units than coarser scales. For example, at the C1 layer, each orientation channel of the finest scale consists of 79 x 79 units, while the coarsest scale has 15 x 15 units. Thus, there are two notions of "template size": the number of C1 units covered by a template, versus the proportion of the image covered by a template. Given a fixed number of C1 units, a template at a coarser scale will cover more of the image than one at a finer scale (Fig. 5.13). Conversely, to represent a given portion of the image, a template at a coarser scale will need to cover fewer C1 units than one at a finer scale. (Note: thus far, we have kept the scale constant (scale 7) and varied the number of C1 units covered)

In Fig. 5.2, we showed that for a particular scale (scale 7 of 9), a large template size (12 x 12 C1 units) results in the CFE, whereas a small template size (4 x 4 C1 units) does not. Thus, it is clear that spatial scale or spatial frequency is not by itself a crucial factor. However, by keeping to a fixed scale, we have confounded number of C1 units with image coverage. We de-confound these next, by keeping the number of C1 units constant, and instead varying the scale.

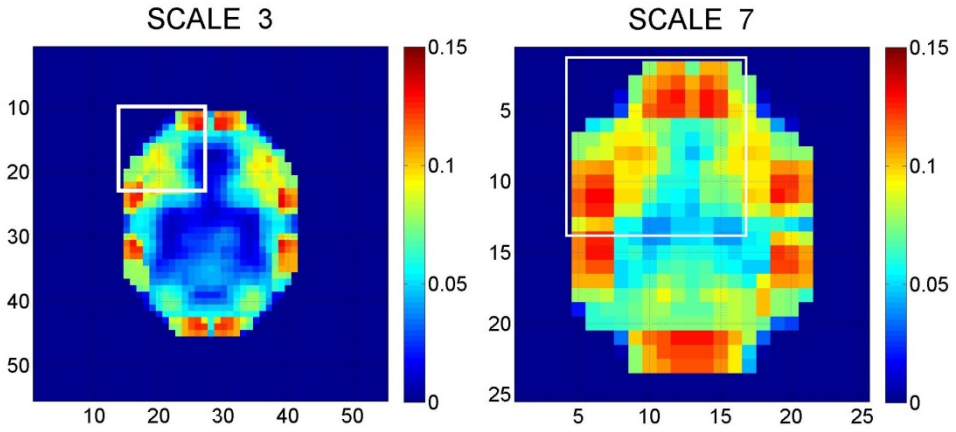


Figure 5.13. Mean C1 activity (averaged over all orientations) at a finer scale (scale 3, left) and a coarser scale (scale 7, right). The entire image or face is represented using more units at the finer scale than at the coarser scale (note the units on axes). Templates covering a particular number of C1 units (e.g. 12x12, white squares), therefore cover more of the image or face for the coarser scale.

5.5.1 Spatial scale

Using a fixed number of C1 units covered by each template (12 x 12), we found that the CFE disappears for finer scales (Fig. 5.14). Thus, the key factor is not the number of C1 units per se, but what proportion of the face is covered by the template.

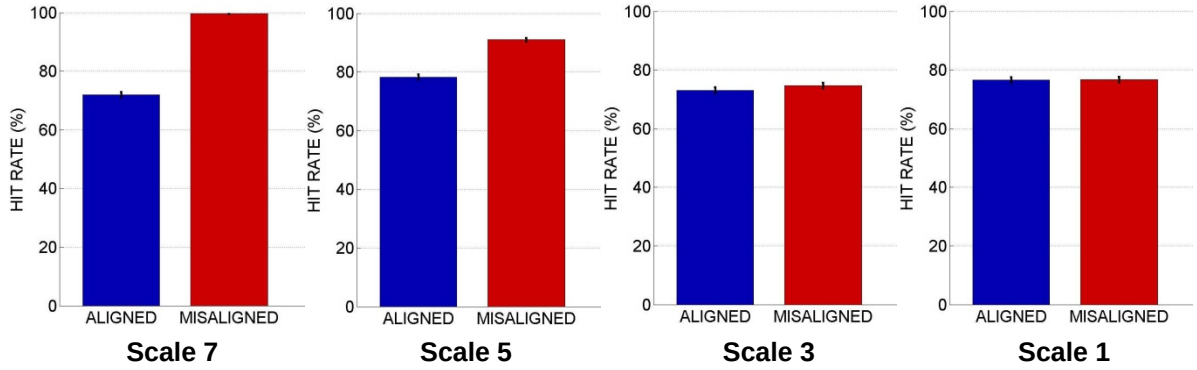


Figure 5.14. Hit-rates for templates extracted from different spatial scales (L to R: scales 7, 5, 3, 1). C1 has 9 scales total, the finest being scale 1. Magnitude of misalignment effect decreases from scales 7 to 5, disappearing for scales 3 and 1.

5.5.2 Image coverage

As additional verification that the key factor is image coverage (not spatial scale) of the templates, we show that even one of the finer scales can also produce the CFE if templates cover enough of the image (Fig. 5.15). This same scale with smaller templates did not show the CFE (results not shown). Note that "image coverage" is not the same as receptive field size. In our model, all these high-level C2 units have the entire image as their receptive fields, since max-pooling is performed over all positions and scales.

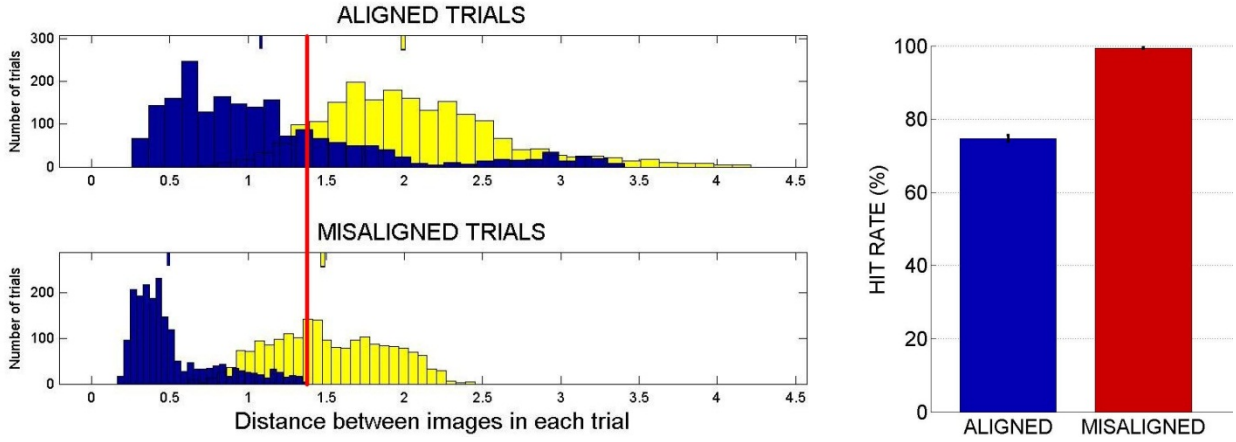


Figure 5.15. Scale 3 (third-finest) can also show a misalignment effect if the templates are sufficiently large in size (24 x 24 C1 units, in this instance).

5.6 Chapter summary

In this chapter, we have established the basic result that our model can produce a misalignment effect (i.e. CFE) by using large, coarse templates. Beyond simply replicating the CFE (e.g. Cottrell et al. 2002), we also show that it is qualitatively robust to various low-level factors.

We made two crucial contributions. Firstly, we provided a step-by-step account of the CFE, linking holism at the single unit level to the misalignment effect at the behavioral level. Secondly, we showed that the key factor in producing holism is the largeness of templates, in terms of image coverage. This sounds unsurprising on hindsight, but has not actually been shown prior to this. Importantly, we show that the key factor is neither spatial scale nor receptive field size per se, factors that can be confused with image coverage.

In the next chapter, we expand our investigation of the CFE and holism by looking at inverted faces.

Chapter 6: CFE for Inverted Faces

Chapter abstract

In this short chapter, we examine the effect of inversion on the CFE. We find that for our model, the CFE is reduced, rather than absent, for inverted faces. This is consistent with our review in Chapter 2, which suggested that the CFE is differential, rather than absolute. At the same time, we find an "inversion anti-effect" for both face-like and object-like processing. This suggests that inversion alone is the wrong control condition, and misalignment should be used. The most important finding of this chapter is that stimulus-related changes alone (rather than different processing style) can account for the reduced CFE for inverted faces. This suggests that inversion does not really "disrupt holism", as commonly thought.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 6: CFE for Inverted Faces

As reviewed in Section 2.3, there are conflicting results regarding the CFE for inverted faces. Even within the "partial" design alone, results are mixed. Some studies find no CFE, while others find a reduced CFE. As discussed, there are many differences between the conflicting studies; one key difference may be the blocking versus intermixing of upright and inverted conditions. Here, we examine the predictions of our model, and see if it can reconcile the conflicting empirical findings. The results presented in this chapter are for large, coarse templates, unless otherwise noted.

6.1 Effect of inversion

Inversion causes a large drop in the average model unit response, as shown in Fig. 6.1 (for aligned composites, a 42% drop from 0.79 to 0.46). More importantly, however, misalignment has less of an effect on inverted than upright faces. From Fig. 6.2, we see that the change in response due to misalignment is larger for upright than for inverted faces. This is true in both absolute terms (Fig. 6.2) and in terms of percentage change (results not shown).

How does this differential effect on model unit responses translate into effects on distances? From Fig. 6.3, we see that misalignment causes little effect for inverted faces, compared to upright faces. In particular, we see that the distances for the "same" trials (blue) decrease significantly (e.g. from the blue hanging bars that indicate the means) for upright faces, but not for inverted faces. For any given threshold (e.g. red line in Fig. 6.3), this decrease in distance translates into an increase in hit rate, i.e. a CFE (a.k.a. "misalignment effect").

Importantly, however, there is a small but noticeable change in the distances for inverted faces (Fig. 6.3 bottom panel). From Fig. 6.4, we see that for inverted faces (red hues), there is a small but distinct CFE. For both upright and inverted faces, the CFE magnitude can vary with threshold, unsurprisingly. As a control, we see that for small, fine features, there is very little change in distance as a result of misalignment (Fig. 6.5), nor is there a CFE (Fig. 6.6).

Interestingly, however, there is an "inversion anti-effect" (i.e. higher hit rate for aligned-inverted than aligned-upright) for both large, coarse and small, fine features. We term this an "anti-effect" because there is a performance improvement, rather than the performance decrement associated with the FIE. The mechanisms for the "inversion anti-effect" are nonetheless very similar to those underlying the "misalignment effect". Since inverted faces elicit smaller responses and distances, for a given threshold, this translates into a higher hit rate. As predicted by our model, this "inversion anti-effect" is found in virtually all "partial" design studies that included inverted faces: Young et al. 1987 (Table 2), Goffaux & Rossion 2006 (Figs. 4 and 5), Robbins & McKone 2007 (Fig. 7), McKone 2008 (Fig. 3), Mondloch & Maurer 2008 (Figs. 2 and 4), Rossion & Boremanse 2008 (Fig. 3) and Soria Bauser et al. 2011 (Fig. 3). Only Carey & Diamond 1994 (Figs. 2, 4, and 5) found opposite results, but blocking of orientation could account for this (e.g. by allowing different strategies or thresholds to be used for upright and inverted faces).

However, if inversion "disrupts holism" like misalignment does, then these results are not surprising for faces. Therefore, we now check if this "inversion anti-effect" is found for non-faces, as predicted from Fig. 6.6. Only two studies used the "partial" design for non-faces. Robbins & McKone 2007 (Fig. 7) found non-significant differences. Soria Bauser et al. 2011 (Figs. 3 and 4) found non-significant differences in two conditions, and results contrary to our prediction in one condition. Therefore, based on the limited evidence so far, our model's prediction does not hold true. Nonetheless, since the regular inversion effect has been found for non-faces (just smaller than for faces), and we because believe that this "inversion anti-effect" stems from the same mechanisms (but manifested differently), we maintain that a more thorough investigation of this issue will validate our prediction.

Overall, our results suggest that inversion may not be the right control to demonstrate face-specificity of the CFE, unless it is used together with misalignment. There are several studies that have inversion without misalignment (e.g. Hole 1994, Hole et al. 1999, Goffaux 2009), and these studies may need to be replicated with misalignment included.

6.2 Reconciling conflicting empirical results

Can our model account for the conflicting empirical results? The CFE is generally small for inverted faces, and the effect is reduced at very high and very low hit rates (Fig. 6.4). This suggests that the conflicting studies may differ quantitatively, rather than qualitatively. Since the CFE for inverted faces is small even under noiseless modeling conditions, various experimental factors such as blocking of conditions or inter-subject variability could make the CFE appear to be absent in some cases (see Section 2.3.1 for further discussion).

In our modeling, we have shown the choice of threshold can affect the magnitude of the CFE. We are not suggesting that this is the only relevant factor, but rather that this factor alone demonstrates the fact that seemingly qualitative differences may in fact be quantitative ones. Crucially, our results suggest that with respect to the issue of upright versus inverted faces, the CFE is a differential (or disproportionate) effect, rather than an absolute one.

6.3 Chapter summary

In this chapter, we have expanded our investigation of the "partial" design CFE by examining the effects of inversion, again providing a step-by-step account of the underlying mechanisms. By showing that the CFE is smaller (rather than absent) for inverted than for upright faces, we have demonstrated that according to our model, the CFE is a differential effect, not an absolute one.

Importantly, since the computational mechanisms were identical for both upright and inverted faces, we have shown that stimulus-related changes alone can account for the reduction of CFE magnitude for inverted faces. Thus, our model suggests that inversion does not "disrupt holism" per se, as is commonly thought.

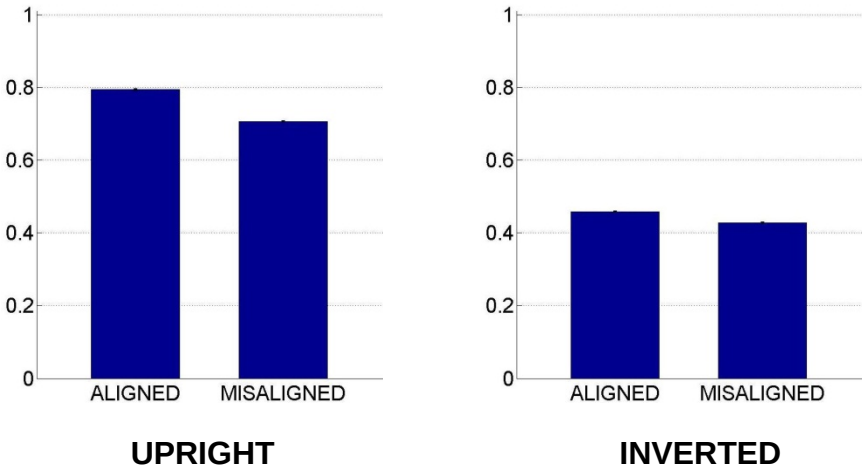


Figure 6.1. Mean responses for large, coarse templates to upright (left) and inverted (right) composites. (SEM error bars are barely noticeable)

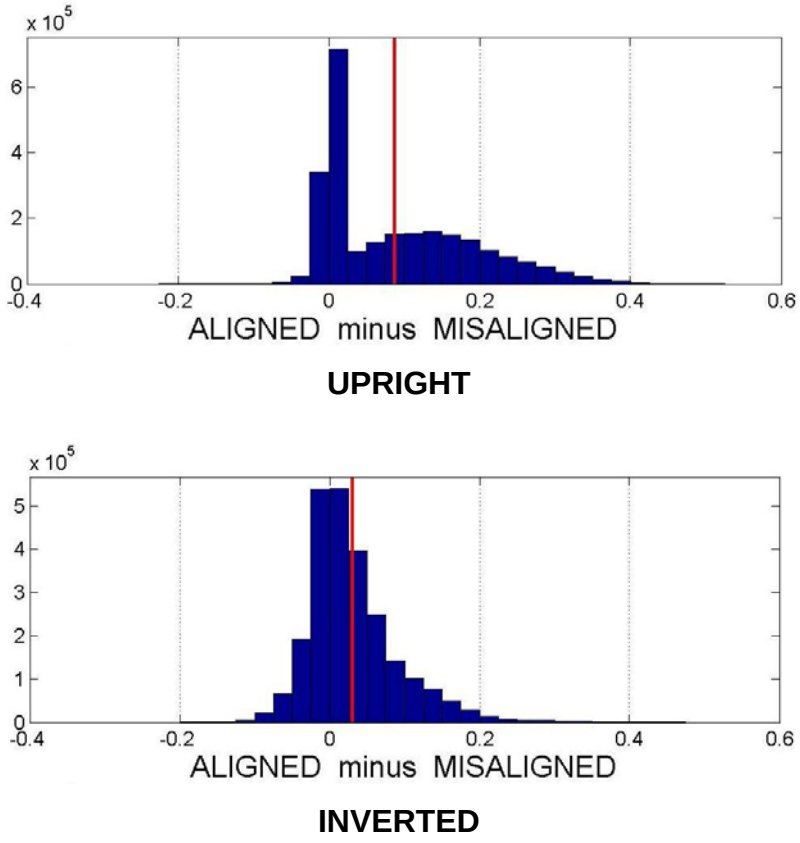
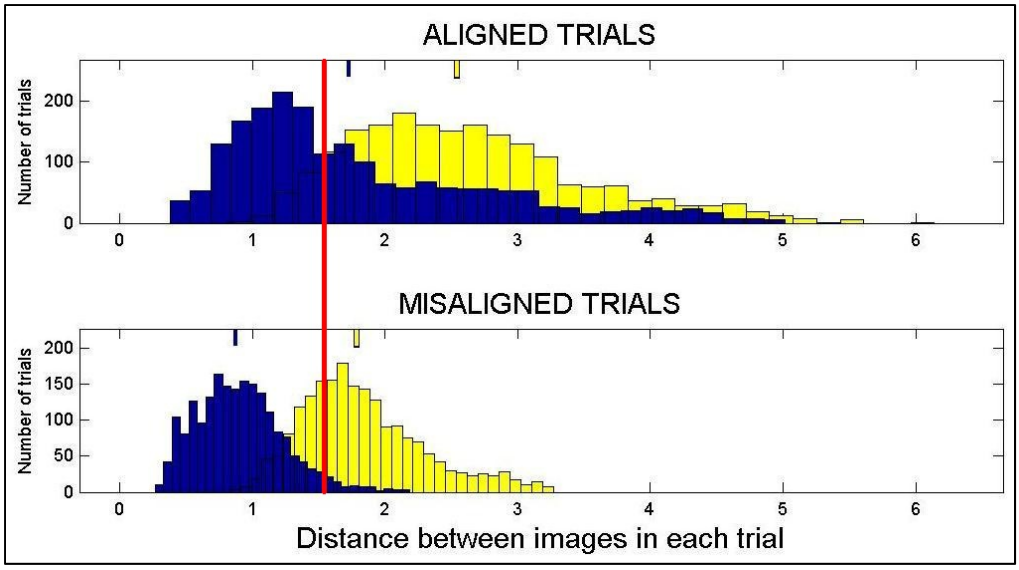
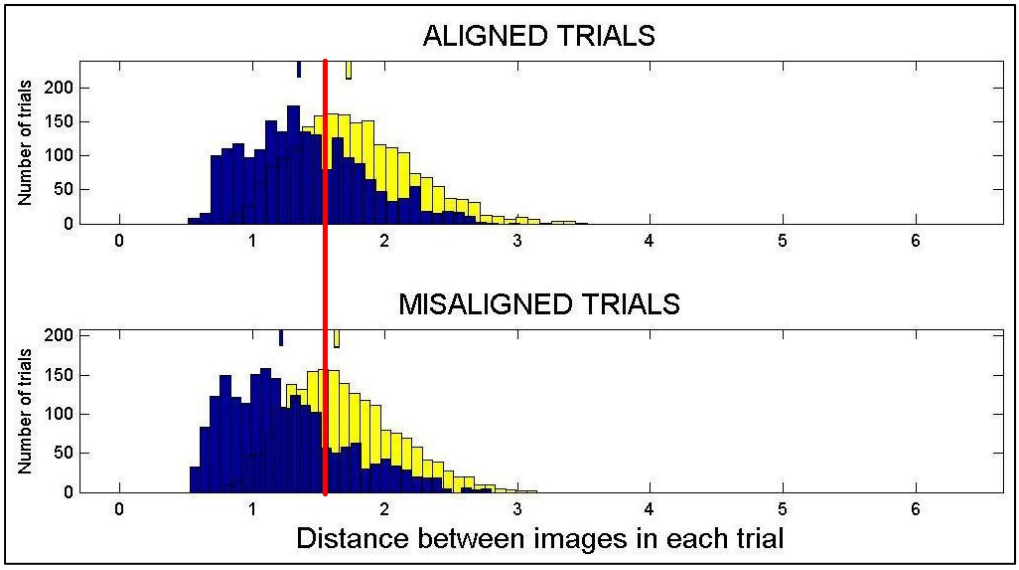


Figure 6.2. Histogram of change in response as a result of misalignment for upright (top) and inverted (bottom) faces. Y-axis: number of model units. Red line indicates mean of distribution.



UPRIGHT



INVERTED

Figure 6.3. Histograms of distances for large, coarse templates for upright (top panel) and inverted (bottom panel) faces. Blue: "same" trials. Yellow: "different" trials. Red line: arbitrary threshold (set to 1.5 in this figure for illustration purposes only). Hanging bar indicates mean of distribution.

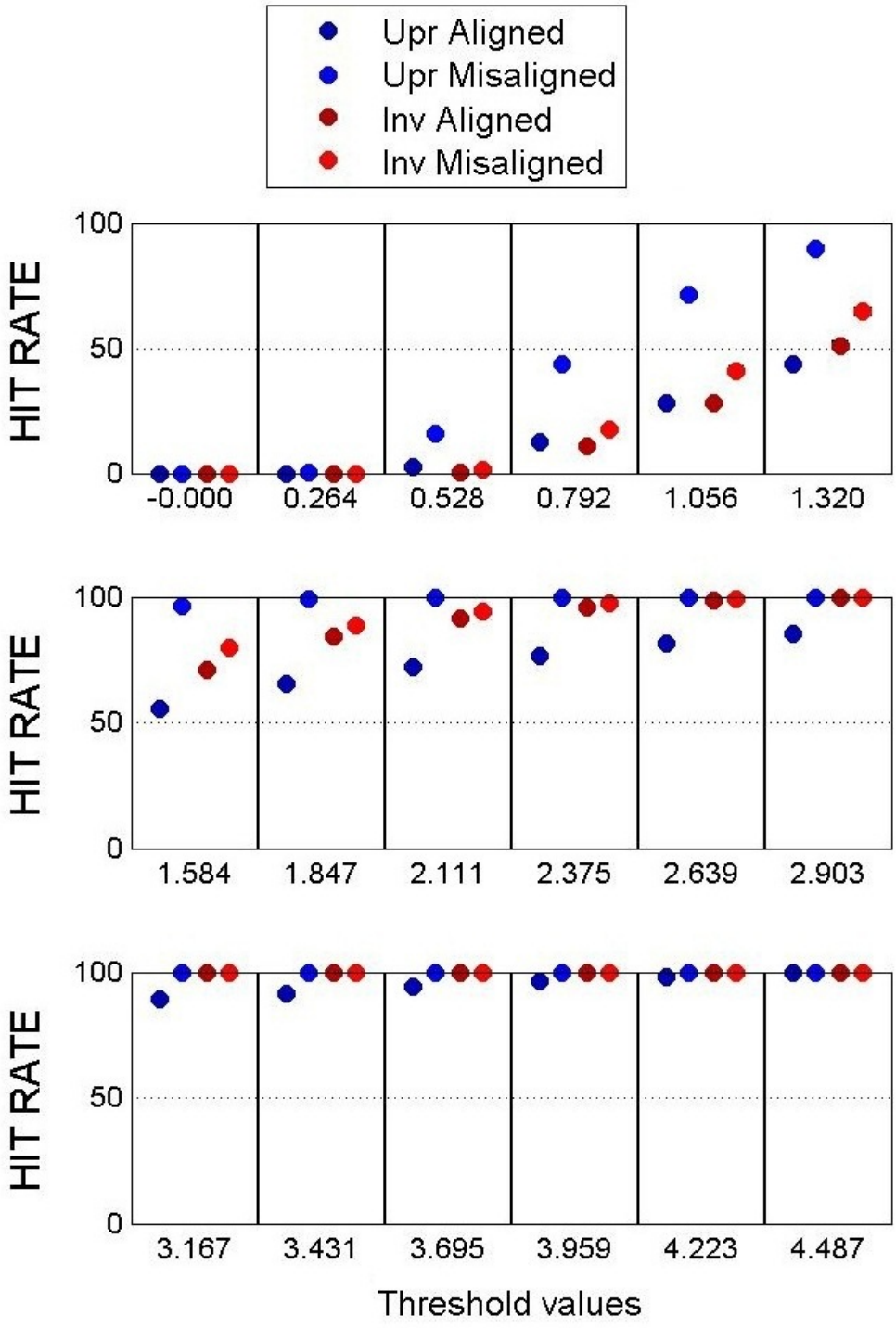
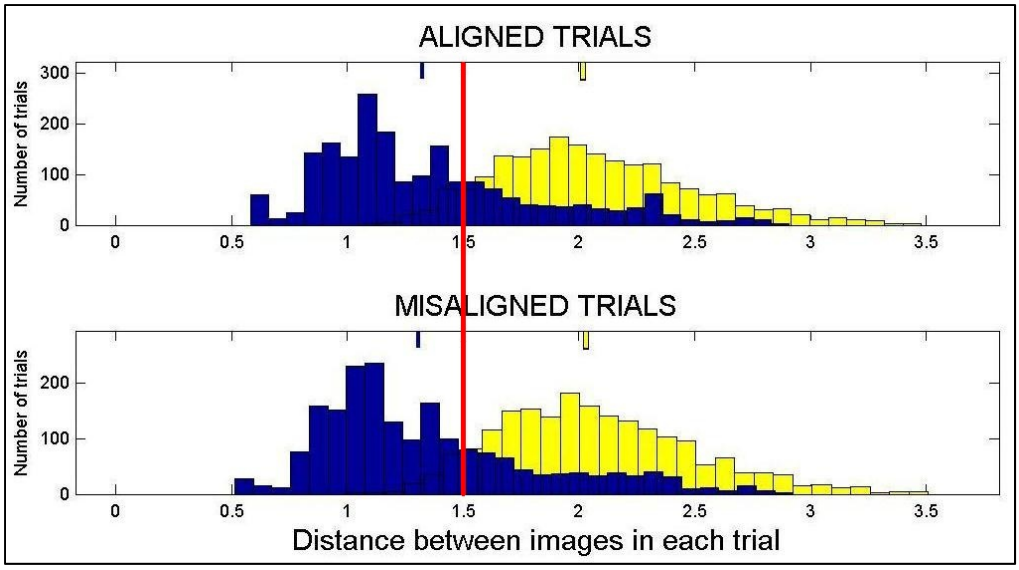
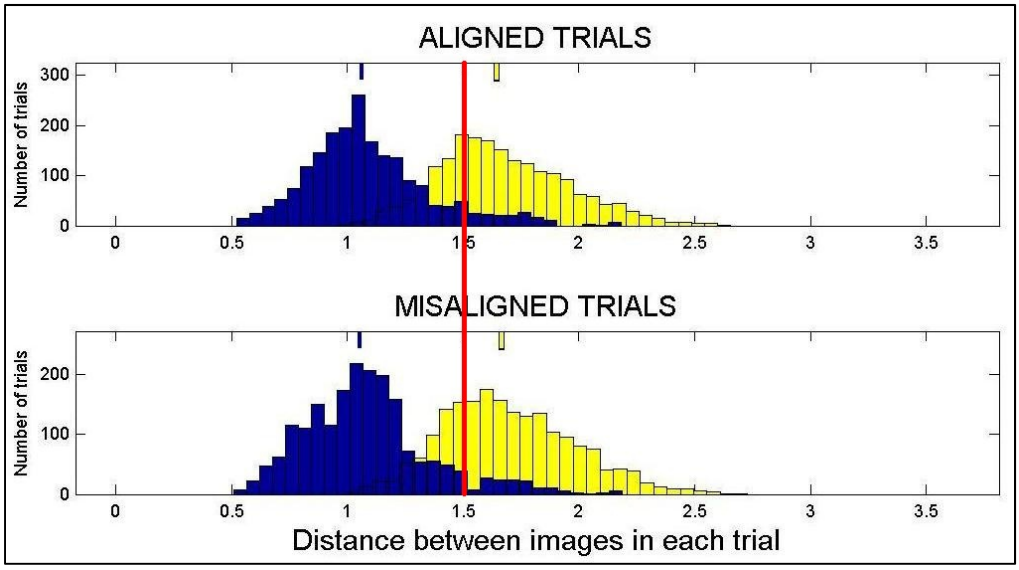


Figure 6.4. Hit rates for large, coarse, features over a wide range of threshold values. Upr: upright. Inv: inverted. SEM error bars are plotted, but are smaller than the dots representing the mean hit rates.



UPRIGHT



INVERTED

Figure 6.5. Histograms of distances for small, fine templates for upright (top panel) and inverted (bottom panel) composites. Blue: "same" trials. Yellow: "different" trials. Red line: arbitrary threshold (set to 1.5 in this figure for illustration purposes only). Hanging bar indicates mean of distribution.

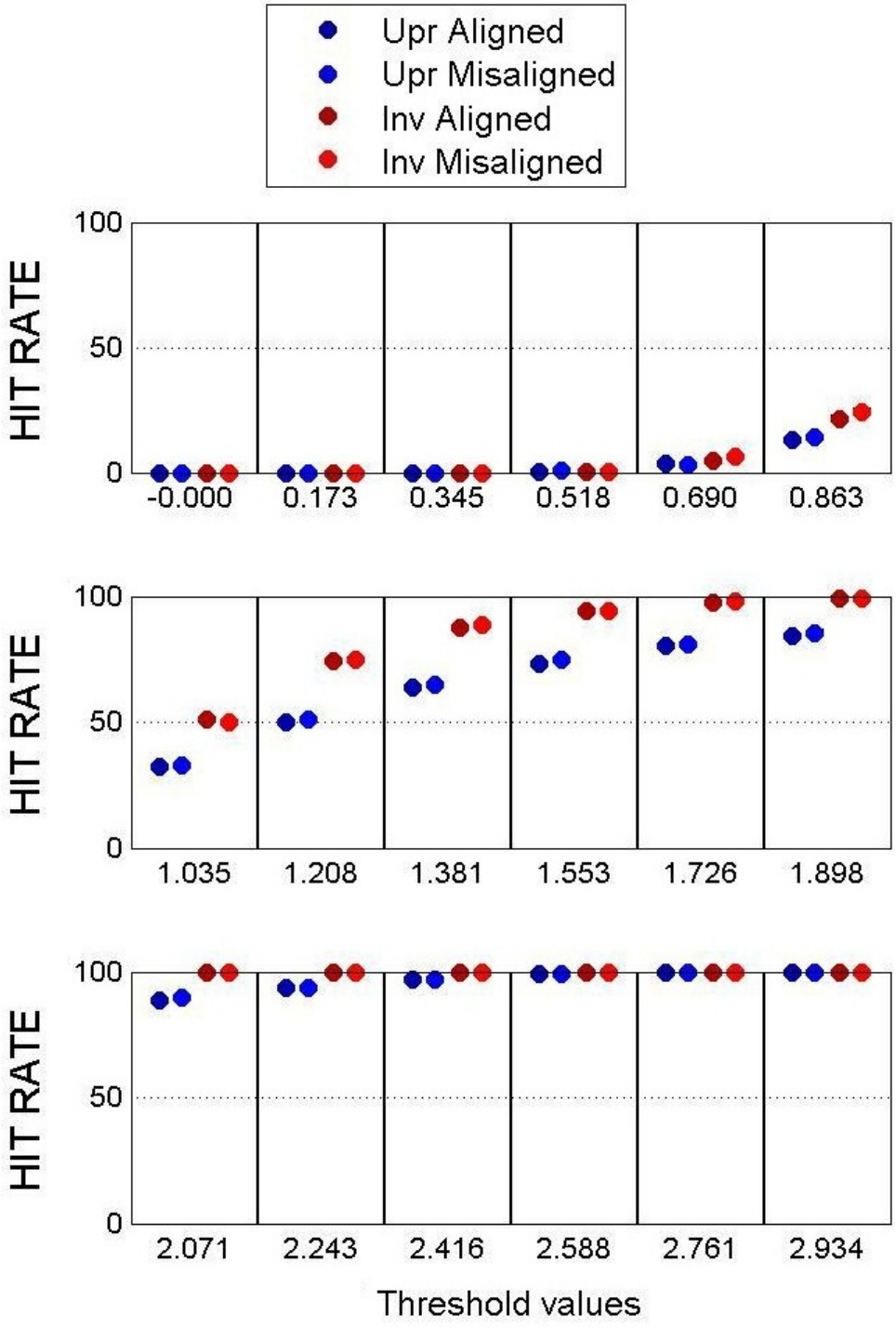


Figure 6.6. Hit rates for small, fine features over a wide range of threshold values. Upr: upright. Inv: inverted. SEM error bars are plotted, but are smaller than the dots representing the mean hit rates.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 7: The "Complete" Design

Chapter abstract

In this chapter, we extend our results by also accounting for the so-called "complete" experimental design for the CFE. In doing so, we show that holistic processing is consistent with both the "partial" and "complete" designs. Importantly, our results provide clarification and justification as to what the right metric in the "complete" design should be.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 7: The "Complete" Design

In Section 2.5, we discussed and compared the two experimental designs for investigating the CFE. To briefly recap, the "complete" design has two additional conditions absent in the "partial" design (See Fig. 7.1), hence the terminology. Whereas holistic processing in the "partial" design is reflected in a "misalignment effect" (higher hit rate for misaligned than aligned trials), in the "complete" design, holistic processing is reflected in a "congruency effect" (higher D' for congruent than incongruent trials). Additionally, the congruency effect is reduced for misaligned or inverted faces, and this is termed a "(congruency x alignment) interaction".

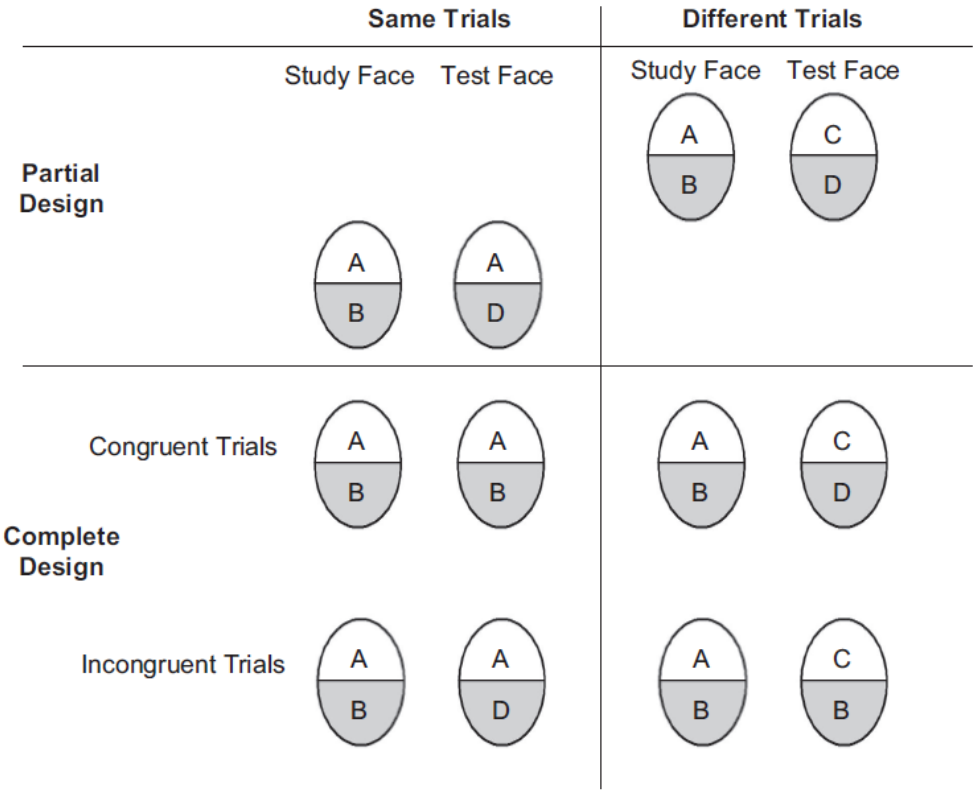


Figure 7.1. Trial types for the CFE "partial" and "complete" designs. Congruent trials are those in which the top and bottom halves are either both same or both different. Note that the "partial" design is a subset of the "complete" design. (Figure reproduced from Cheung et al. 2008. See p.16 for copyright notice.)

In this section, our primary goal is to show that our model can also replicate the CFE for the "complete" design. As discussed in Section 2.2, the logic underlying both designs is essentially identical. Therefore, a model of holism should be expected to show the CFE using either design.

Note that here, we vary only the experimental design, keeping everything else identical. This is so that we can be sure that any differences are solely due to experimental design. This is unlike the empirical studies that have been conducted, where differences in experimental design are confounded with procedural differences (discussed in Section 2.5.5). In particular, here, both first and second composites are either aligned or misaligned. Also, the attentional modulation is identical for both first second composites.

7.1 Results

From Fig. 7.2, we see that our model does indeed reproduce both characteristics of the CFE for the "complete" design. For aligned composites, D' is higher in the congruent than incongruent condition (i.e. there is a congruency effect). This congruency effect is reduced or absent for misaligned composites (i.e. there is a congruency x alignment interaction).

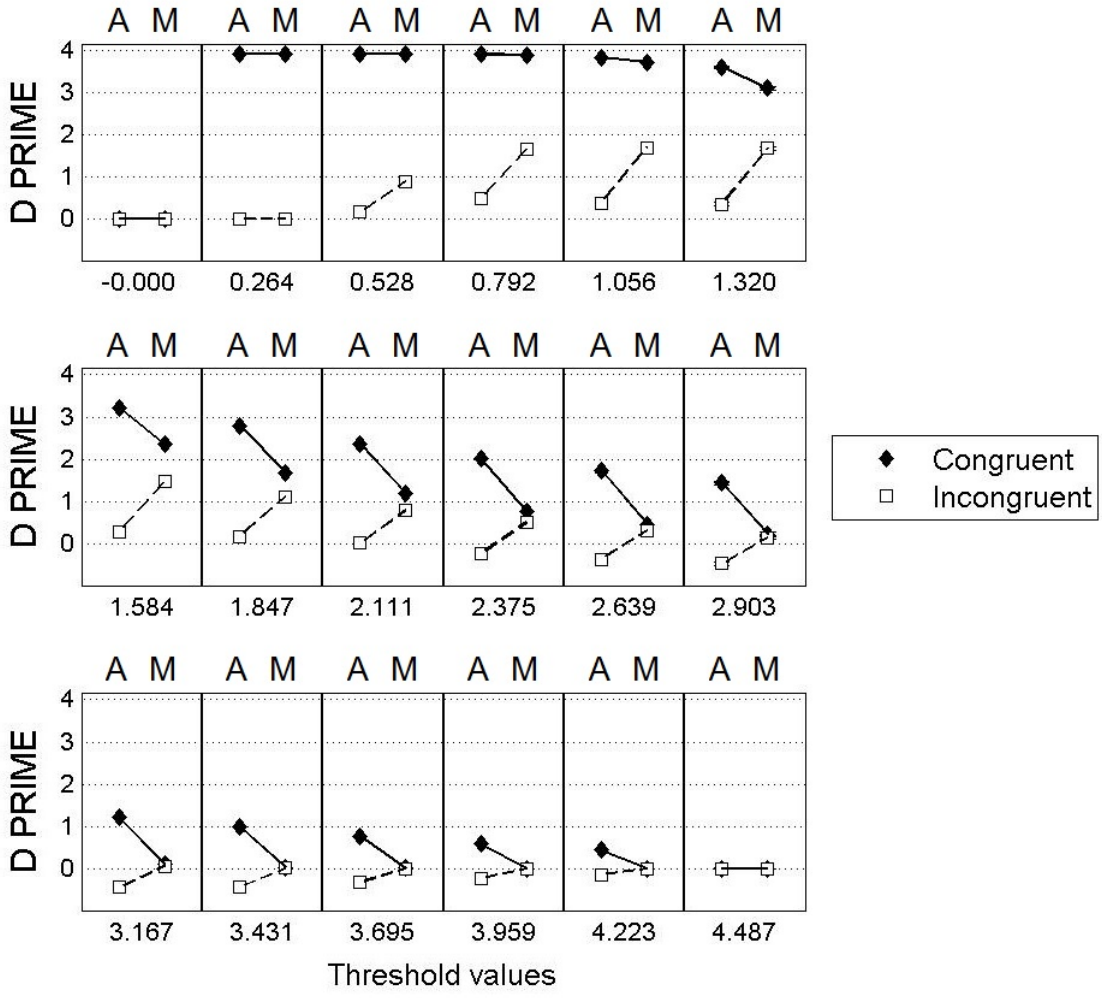


Figure 7.2. D' results in the "complete" design for large, coarse features, shown for the full range of thresholds. A: aligned. M: misaligned.

7.1.1 Detailed explanation for large, coarse features

In order to gain a mechanistic understanding of the congruency effect, we now examine in detail the distances between composites for each of the four conditions, as well as the effect of misalignment on each of these. The distribution of distances is shown in Fig. 7.3

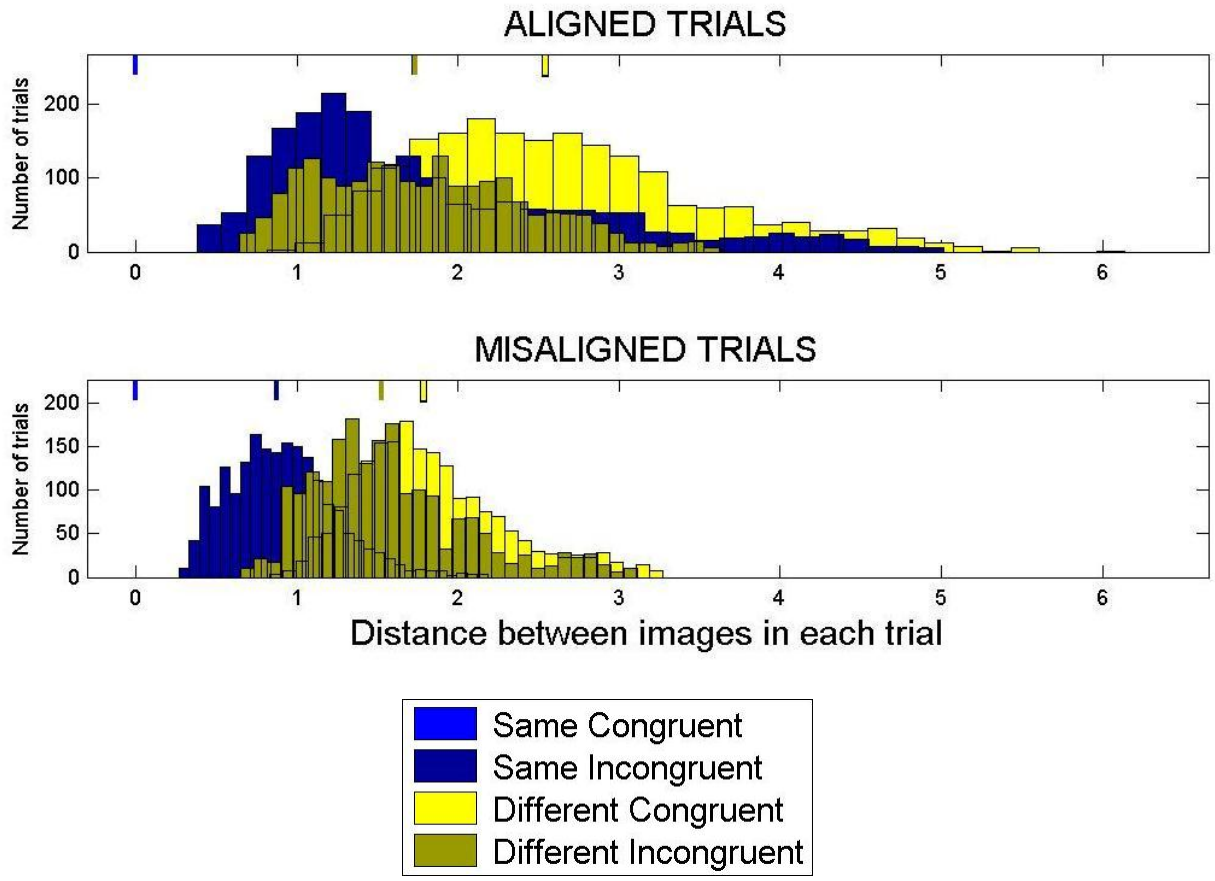


Figure 7.3. Histograms of distances for large, coarse features. Top: aligned trials. Bottom: misaligned trials. Blue hues: "same". Yellow hues: "different". Darker shades: incongruent. Brighter shades: congruent. Hanging bar indicates mean of distribution. Note: for aligned trials, dark yellow hanging bar obscures dark blue.

First, we concentrate on the aligned trials (Fig. 7.3 top). For the same-congruent condition, the distances are all 0 (indicated by bright blue hanging bar only). This is not surprising, since the first and second composites are literally identical, and our model is noiseless. For the same-incongruent condition (dark blue), distances are somewhat similar to the different-incongruent condition (dark yellow), hence the very low D' in the incongruent condition (Fig. 7.2). Importantly, the distance separation between the congruent trials (bright blue and bright yellow) is larger than that between the incongruent trials (dark blue and dark yellow) – hence the congruency effect.

As predicted in Section 2.2, usage of D' is not necessary to see a congruency effect. For any given threshold, congruent-same trials (bright blue; distance 0) will have a higher hit-rate than incongruent-same trials (dark blue). At the same time, congruent-different trials (bright yellow) will have lower false-alarm rate than incongruent-different trials (dark yellow). In other words, separately for both "same" and "different' trials, congruent trials have better performance than incongruent trials.

When composites are misaligned (Fig. 7.3 bottom), distances generally become smaller. Crucially, however, the separation between incongruent-same (dark blue) and incongruent-different (dark yellow) conditions is increased (compared to aligned composites), leading to a higher D' (see Fig. 7.2). At the same time, the separation for congruent trials (bright blue and bright yellow) is decreased compared to aligned composites, leading to a lower D'. Together, these result in a smaller congruency effect for misaligned than aligned composites.

Results using large but fine features are qualitatively the same (results not shown) as when using large, coarse features.

7.1.2 Results for small, fine features

We next examine the results using small, fine ("object-like") features. Somewhat surprisingly, a congruency effect is also found (see Fig. 7.4). Importantly, however, misalignment does not affect this congruency effect. In other words, there is no (congruency x alignment) interaction.

In Section 2.5.5, we examined the congruency effect for various "complete" design studies and hypothesized that a congruency effect was found even in object novices primarily due to "contextual induction". This is not the case here, however; both composites receive attentional modulation. So why do our "object-like" features show a congruency effect?

In our model, attentional modulation is not absolute. Therefore, the bottom halves are still "perceived" to some extent (and exacerbated by the use of *ndp*; see Section 5.4.6). Therefore, even for our "object-like" features, congruency (agreement between top and bottom halves) still matters.

What does this mean in practice, for empirical studies? We do not make any claims about the realism of our attentional modulation scheme. However, attentional modulation in human subjects is unlikely to be absolute either (for both objects and faces). Thus, our modeling results suggest that in additional to contextually-induced "holism", another reason why congruency effects can be found for objects in novices is because the bottom halves are still at least partially attended and processed.

More importantly, our results suggest that there are two somewhat separable contributions to the empirical measurement of holism. The first is the "holism" of the features. In our model, this corresponds to the proportion of a whole face that each feature is tuned to. The second is how successfully modulated the processing of the to-be-ignored bottom halves is. This is not traditionally what is considered to be "holism", since it is more related to attentional mechanisms

that are not face-specific. These two contributions can be construed as "perceptual integration" versus "selective attention" (discussed in Section 2.8). However, these two things may not be completely separable, and more thought needs to be given as to how best to measure "perceptual integration" per se.

In any case, our modeling results provide justification for why the signature of holistic processing in the "complete" design should be a (congruency x alignment) interaction, rather than a congruency effect per se (as discussed in Section 2.5.2). Specifically, because of imperfect attentional modulation mechanisms, even object-like processing will produce congruency effects.

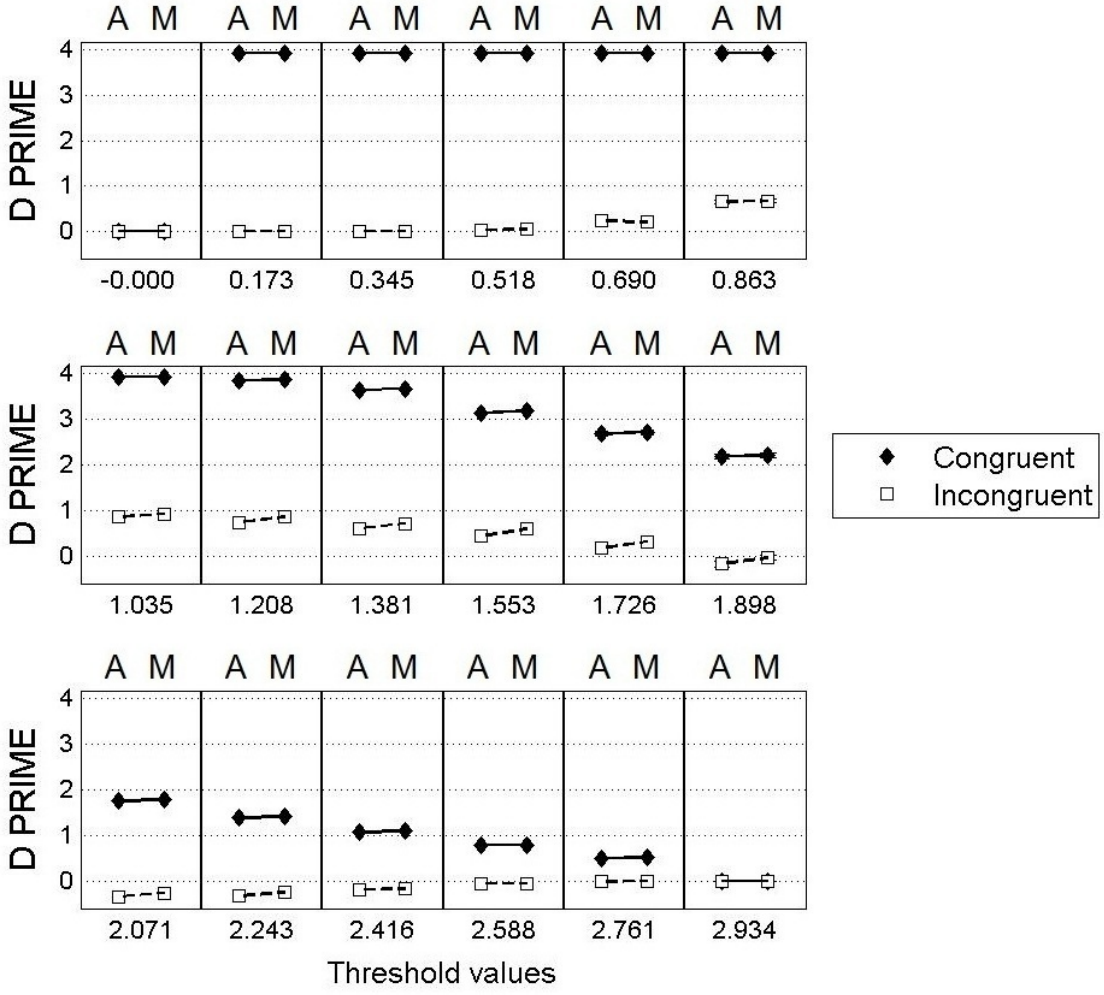


Figure 7.4. D' results in the "complete" design for small, fine features, shown for the full range of thresholds. A: aligned. M: misaligned.

Does this mean that the "partial" design is better at isolating "perceptual integration" than the "complete" design? No. As we can see from Fig. 7.5, for small, fine features, all four conditions (same/different x congruent/incongruent) are equally resistant to misalignment (unlike in Fig. 7.3). In other words, as we have emphasized, both designs are equally valid. The issue with the "complete" design was the wrong metric (congruency effect), possibly due to incomplete reasoning (e.g. reasoning about how holism produces congruency effect, but neglecting other factors that may also contribute to producing a congruency effect). This is precisely one of the advantages of quantitative modeling. It has aided us in answering the question: "holistic processing causes a congruency effect, but does the presence of a congruency effect necessarily imply that processing is holistic?" We have shown that the answer is no.

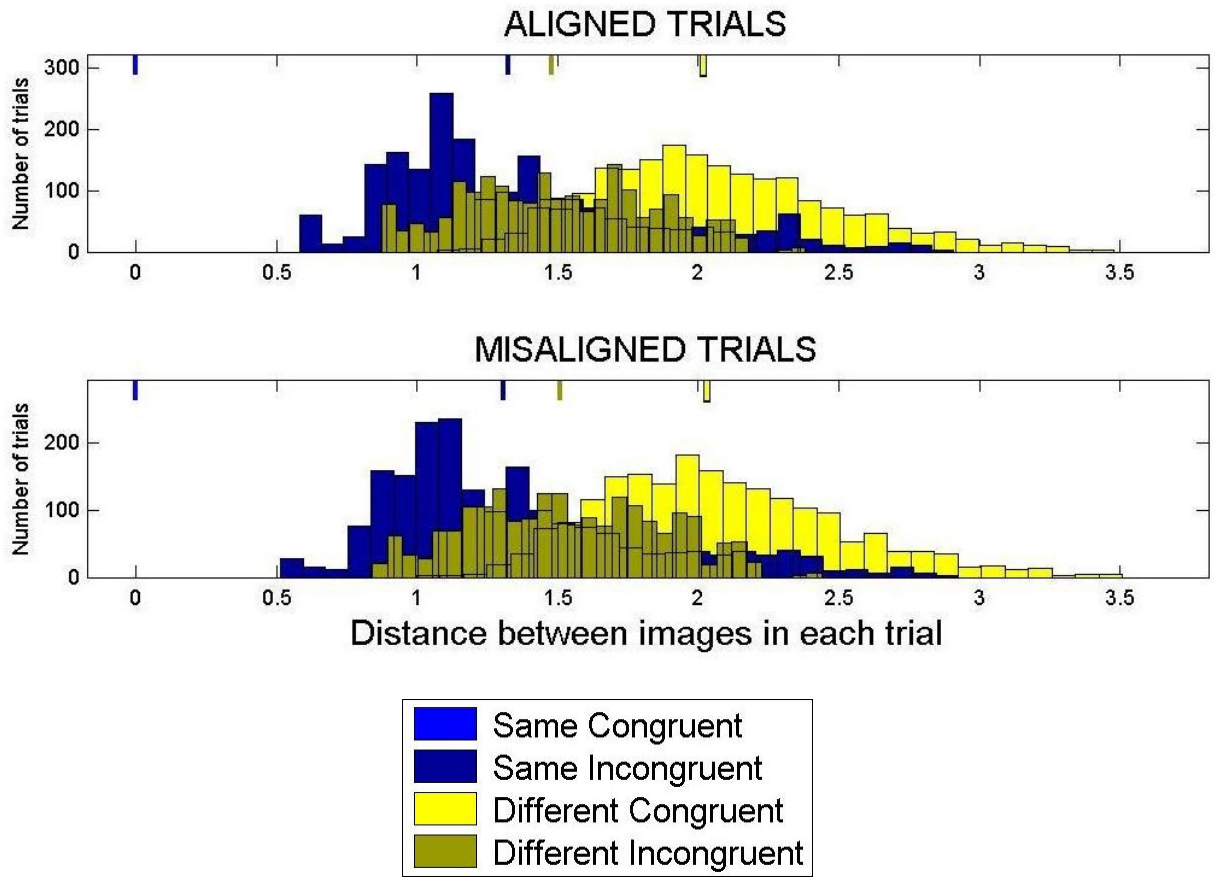


Figure 7.5. Histograms of distances for small, fine features. Top: aligned trials. Bottom: misaligned trials. Blue hues: "same". Yellow hues: "different". Darker shades: incongruent. Brighter shades: congruent. Hanging bar indicates mean of distribution.

7.2 Which is the better paradigm?

Based on our modeling results, we confirm our intuition that both designs are equally valid (if the right metric is used). Nonetheless, it is important to note some practical issues. We have

discussed these in Section 2.5.7, but reiterate them here. Analysis in the "complete" design for the (congruency x alignment) interaction uses four times as many trials as in the "partial" design. If this is equated, both designs are likely to be equally robust.

However, there is an important caveat, as highlighted by Gauthier and colleagues. The "partial" design is more susceptible to fluctuations in threshold, i.e. "bias" (in the sense of inexplicable propensities to favor "same" over "different" or vice-versa). This is not necessarily the same as "bias" (a.k.a. "criterion") in the Signal Detection Theory (SDT) sense. Changes in SDT criterion/bias can be due to changes in threshold or changes in signal distribution. For a fixed threshold, changes in SDT criterion/bias can be qualitatively accounted for, simply by the decrease in distances (due to inversion or misalignment, for instance). Instead, the problem arises when conditions (e.g. upright versus inverted) are blocked, not intermixed, or when different populations (e.g. children versus adults) are compared. There, the thresholds may differ in uncontrollable ways, and looking at only the hit-rate may be misleading.

7.3 Chapter summary

In this chapter, we accounted for the "complete" design. In doing so, we showed that holistic processing is consistent with both the "partial" and "complete" designs. Importantly, our results provide justification for what the right metric in the "complete" design should be.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 8: Contrast Reversal

Chapter abstract

Contrast reversal impairs the recognition of faces, but not objects. One might infer from this that contrast reversal disrupts face-specific processing, much like inversion does. However, the CFE is found for contrast-reversed faces. In this chapter, we use our model to reconcile these apparent contradictions. In addition, the model makes several counter-intuitive post-dictions and predictions. Overall, this chapter highlights the importance of a mechanistic, step-by-step understanding.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 8: Contrast Reversal

Another hallmark of face processing is the sensitivity to contrast polarity (Sinha et al. 2006). Faces are much harder to recognize when their contrast is reversed (e.g. Galper 1970, Hayes et al. 1986, Kemp et al. 1996). However, this is not the case for objects (e.g. Gauthier et al. 1998, Nederhouser et al. 2007). Surprisingly, however, the CFE has been found for contrast-reversed (a.k.a "negative") faces (Hole et al. 1999, Calder & Jansen 2005, Taubert & Alais 2011). If contrast reversal impairs face processing (possibly like misalignment or inversion), then why are contrast-reversed faces nonetheless processed holistically?

Can our model "predict" (post-dict) and reconcile the body of results that have been found? Importantly, we examine our model "as-is", without any further modifications to fit the existing literature regarding contrast reversal. To do so, we delve into the sometimes puzzling (even contradictory) findings that have been uncovered since the first study by Galper in 1970.

8.1 Mini-review of contrast reversal

We do not aim to provide a comprehensive review of contrast reversal in this section. Rather, we will give a broad sampling of key results. In particular, we do not shy away from presenting conflicting studies; one of our goals here is to see if our model can reconcile these.

The most basic result of sensitivity of face processing to contrast reversal is a robust one. It has been found for identification of famous faces (Gilad et al. 2009), and for novel faces in seen/unseen (Liu & Chaudhuri 1997) and two-alternative forced-choice (Galper 1970) memory tasks. It has also been found for tasks that minimize confounds from memory-related effects: simultaneous match-to-sample (Nederhouser et al. 2007) and simultaneous same/different (Robbins & McKone 2007) tasks. In contrast, this sensitivity has not been found for "Greebles" (Gauthier et al. 1998), chairs (Subramaniam & Biederman 1997), and "blobs" (Nederhouser et al. 2007).

Subtle distinctions must be made, however. It is clear that when study/test or sample/match faces are of opposite contrast (positive/negative or negative/positive, a.k.a. PN and NP respectively), performance is much worse than in the positive/positive (PP) condition. However, results are mixed for a fourth condition: negative/negative (NN).

Two studies found that face discrimination performance is significantly worse in the NN condition than in the PP condition (Russell et al. 2006, Robbins & McKone 2007). Interestingly, this difference was also found (albeit reduced) for dog stimuli in novices (Robbins & McKone 2007). However, Liu & Chaudhuri (1997) found that performance in the NN and PP conditions did not differ significantly. In addition, performance in both the NN and PP conditions were significantly better than in the NP and PN conditions (which did not differ significantly).

On the other hand, there is unanimous agreement among three studies that examined holistic face processing using the CFE in both NN and PP conditions. All studies used the "partial" design.

Conditions were either blocked (Hole et al. 1999, Calder & Jansen 2005) or intermixed (Taubert & Alais 2011). Composites were presented either simultaneously (Hole et al. 1999, Calder & Jansen 2005) or sequentially (Taubert & Alais 2011). The controls consisted of either inverted (Hole et al. 1999) or misaligned faces (Calder & Jansen 2005, Taubert & Alais 2011). Holistic processing was found for both identity (Hole et al. 1999, Taubert & Alais 2011) and expression (Calder & Jansen 2005). Altogether, these studies show that the findings are generally robust, and not specific to any particular experimental variables.

Few electrophysiological studies have been conducted to examine the effects of contrast reversal. To the best of our knowledge, only one published study has investigated this (but see Ohayon et al. 2010 for an unpublished study). Rolls & Baylis (1986) recorded from face-selective neurons in the anterior dorsal part of the superior temporal sulcus (STS) in area TPO and on the ventral lip of the sulcus in areas TEm and TEa. Testing 42 neurons with positive- and negative-contrast faces, they found that as a population, these neurons responded essentially equally strongly to both sets of faces, since the slope of the regression line is 0.9 (see Fig. 8.1). How these data fit with the behavioral results is a further mystery, but our model will shed some light on this.

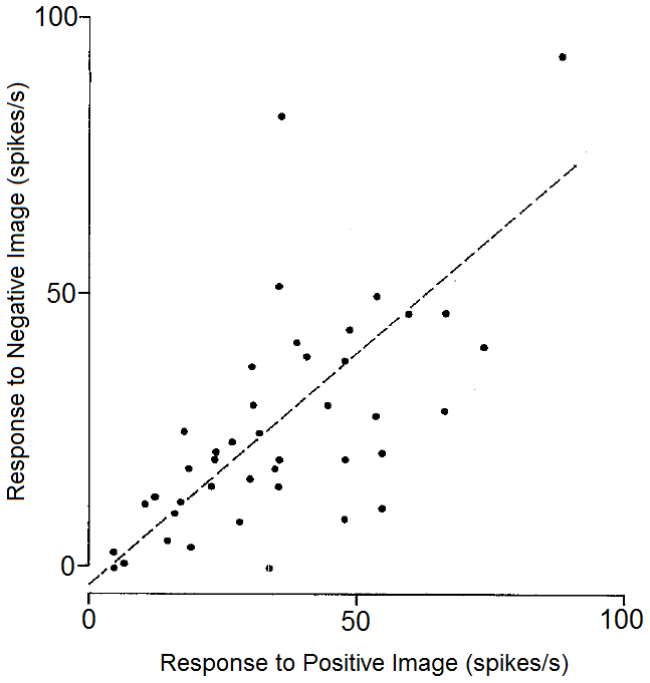


Figure 8.1. Effect of contrast reversal on neuronal response. Each point represents one neuron. Slope of the regression line (dashed) is 0.9. Figure reproduced from Rolls & Baylis (1986). See p.16 for copyright notice.

In summary, there are four sets of findings that we would like our model to account for:

1) The CFE was found for contrast-reversed faces
2) Face recognition performance for NP and PN is worse than for PP
3) Object recognition performance for NP and PN is as good as for PP
4) Conflicting results for face recognition performance for NN versus PP

We will first show that the model is able to reproduce the CFE for contrast reversed faces (Finding #1). Then, we examine in detail the effect of reversal at the single unit, population and behavioral levels in our model. We will then use this detailed understanding of contrast reversal effects in our model to account for the patterns of recognition performance (Findings #2 and #3), and attempt to account for the conflicting results in Finding #4.

8.2 CFE for contrast-reversed faces

Without any modification to our model, we ran it on contrast-reversed faces to see if it reproduced the CFE. Contrast reversal was performed at the pixel level (Fig. 8.2) by computing for each pixel the number $255 - x$, where x is the pixel value in the positive-contrast image; 255 (white) is the maximum pixel value. Background pixels were maintained at value 0 (black). All other experimental conditions were unchanged. In particular, the same set of composite faces (but contrast-reversed) was run through the model (with bottom halves attenuated like before), and the same set of large, coarse face templates was used.

Consistent with the reported behavioral findings, the model showed a CFE for contrast-reversed faces. In particular, the hit rates were higher for misaligned (Calder & Jansen 2005, Taubert & Alais 2011) and inverted (Hole et al. 1999) composites than for aligned composites (see Fig 8.3).

While all the existing empirical studies used the "partial" design, we also found that the "complete" design produced a CFE for contrast-reversed faces (Fig. 8.4). This was both in terms of a congruency effect, as well as a (congruency x alignment) interaction. These results are unsurprising given the model's behavior in the "partial" design, but are nonetheless genuine predictions (rather than post-dictions). In light of the continuing debate over the "partial" versus "complete" designs (see Section 2.5), these predictions may be worth testing.

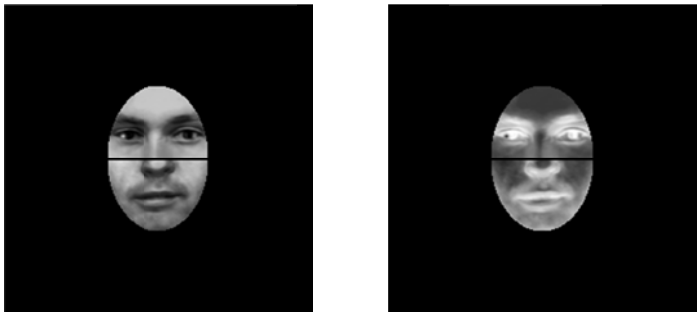


Figure 8.2. Regular (left) and contrast-reversed (right) versions of a typical composite face. The background is not reversed.

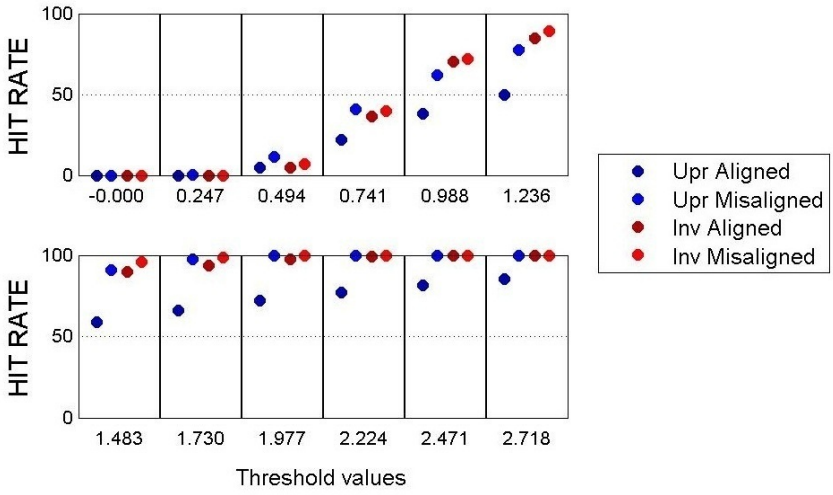


Figure 8.3. Contrast-reversed faces elicited a "misalignment effect" for large, coarse features for virtually all thresholds. Upr: upright. Inv: inverted.

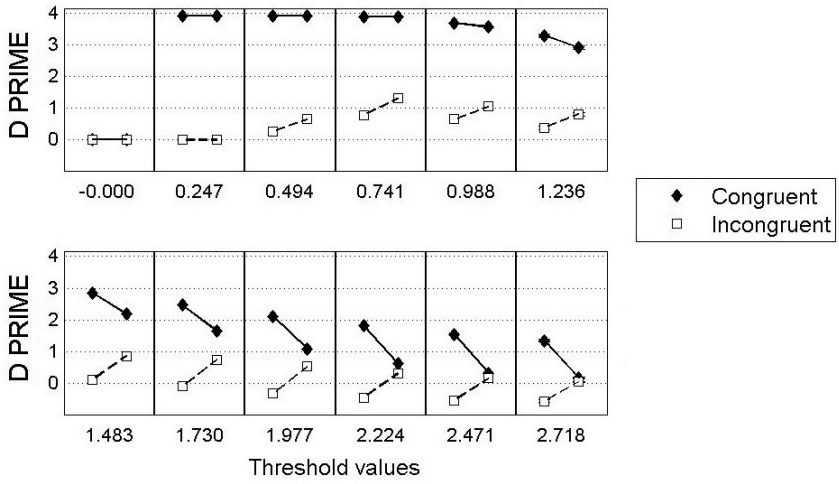


Figure 8.4. Contrast-reversed faces elicited a congruency effect and (congruency x alignment) interaction for large, coarse features for virtually all thresholds.

We find further similarities between results for positive and negative contrast faces. In both cases, the CFE is not dependent on spatial scale per se (i.e. CFE was also found using large, fine features; results not shown here). Also, using "object-like" small, fine features, there is no misalignment effect (Fig. 8.9), but there is an "inversion anti-effect" (see Section 6.1).

We note one interesting difference between the results for positive and negative faces. For positive-contrast faces, misalignment produces a greater boost in hit rate than inversion. For negative-contrast faces, however, both produce roughly equal boosts in hit rate (Fig. 8.3). Further

investigation is required to see if this is a genuine effect in the model, but this may potentially lead to an interesting prediction.

8.3 Step-by-step account

Like we did for regular (positive-contrast) faces in Chapter 5, we take a step-by-step approach to understand why negative faces also elicit a CFE. As we will show, this is the key to reconciling the ostensibly puzzling results for contrast reversal in recognition tasks.

8.3.1 Effect of contrast reversal on responses

From Fig. 8.5, we can see that unlike misalignment or inversion, contrast reversal does not decrease the mean response much (from 0.79 to 0.73). Furthermore, the responses to positive and negative composites are highly correlated (slope of regression line is 0.91, similar to that in Fig. 8.1). Note, however, that for any particular unit, the response to positive and negative versions of a composite can be quite different (note that many points are relatively far away from the diagonal line of equality).

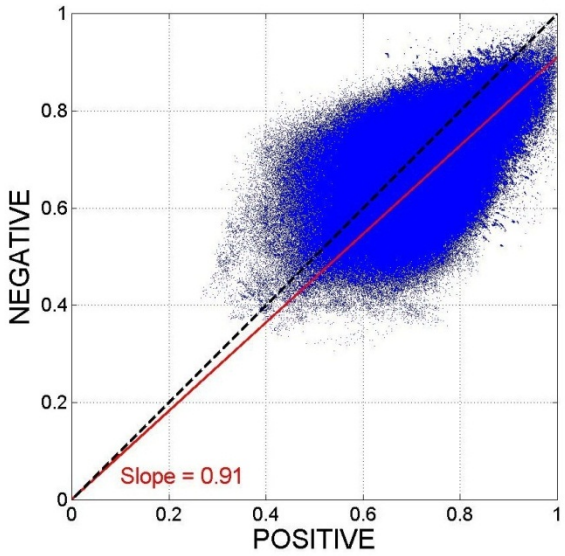


Figure 8.5. Scatter-plot of responses to negative versus positive composites, for large, coarse features. Dashed black line: equal response to positive and negative composites. Red line: regression line constrained to pass through origin.

Why does contrast reversal not have a deleterious effect on the mean response? To understand this, we look at the V1-like responses at Scale 7 (the scale of the large, coarse templates). From Fig. 8.6, we can see that the responses are not drastically different – the eye, nose and mouth regions still produce strong responses. This is not surprising. Complex cells in V1 are strongly invariant to 180 degree phase-shifts (i.e. contrast reversal) in grating stimuli (De Valois et al.

1982, Skottun et al. 1991). Faces are not gratings, hence the responses are less invariant, but similar principles apply. The significance of the model not showing strong invariance will become apparent later (in Section 8.4), when we attempt to account for the effect on contrast reversal on face recognition.

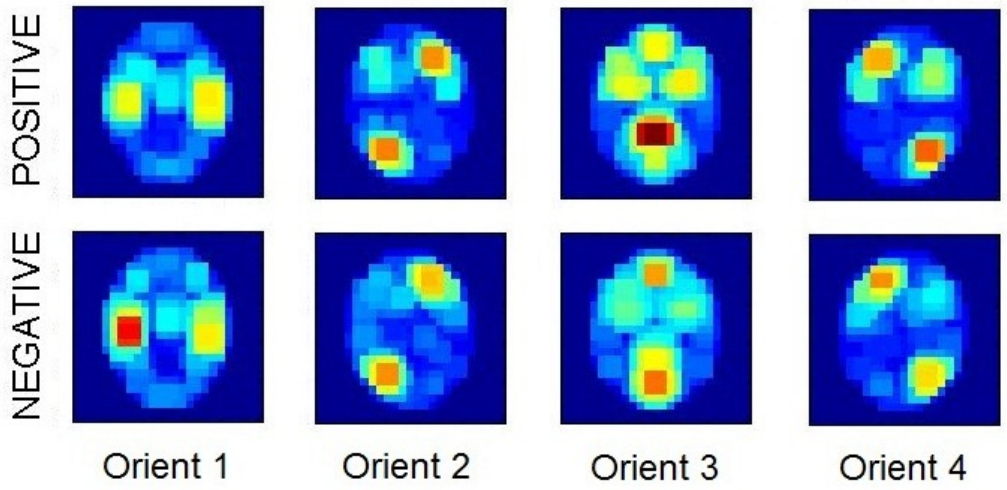


Figure 8.6. C1 responses to positive (top row) and negative (bottom row) versions of a typical composite. Blue: low responses. Red: high responses. Orientation 1: vertical. Orientation 3: horizontal.

Note that perfect invariance to contrast reversal could in theory be achieved. During the V1 simple-cell-like template matching ("S1") stage, if the image patches are normalized to have a mean value of zero (not currently done) and the outputs of the template matching are rectified (currently done), then contrast reversal will have absolutely no effect – in theory. In practice, because we keep the background black instead of also reversing its contrast, V1-like responses at the face boundary regions will be affected by reversal. If the background was also reversed, or was always kept at mid-gray, then perfect invariance to contrast reversal can be achieved in practice.

The previous paragraph might seem overly concerned with ostensibly unimportant details, but it illustrates how such details may potentially have important consequences. For example, Rolls & Baylis (1986) found that contrast reversal had little overall effect on neural responses (Fig. 8.1), but Ohayon et al. (2010, as yet unpublished) found that firing rate was reduced by 50% on average. It is currently unclear why different results were found. Empirical studies have understandably not investigated such issues in detail, e.g. by assuming that the background luminance is unimportant. However, quantitative modeling forces such assumptions to be made explicit and examined.

After looking at the effect of contrast reversal per se, we now turn to the effect of misalignment on positive and negative contrast faces (Fig. 8.7). Because contrast reversal generally did not

reduce responses by much, we see that misalignment has very similar effects on negative (Fig. 8.7 left) and positive (Fig. 8.7 right) faces.

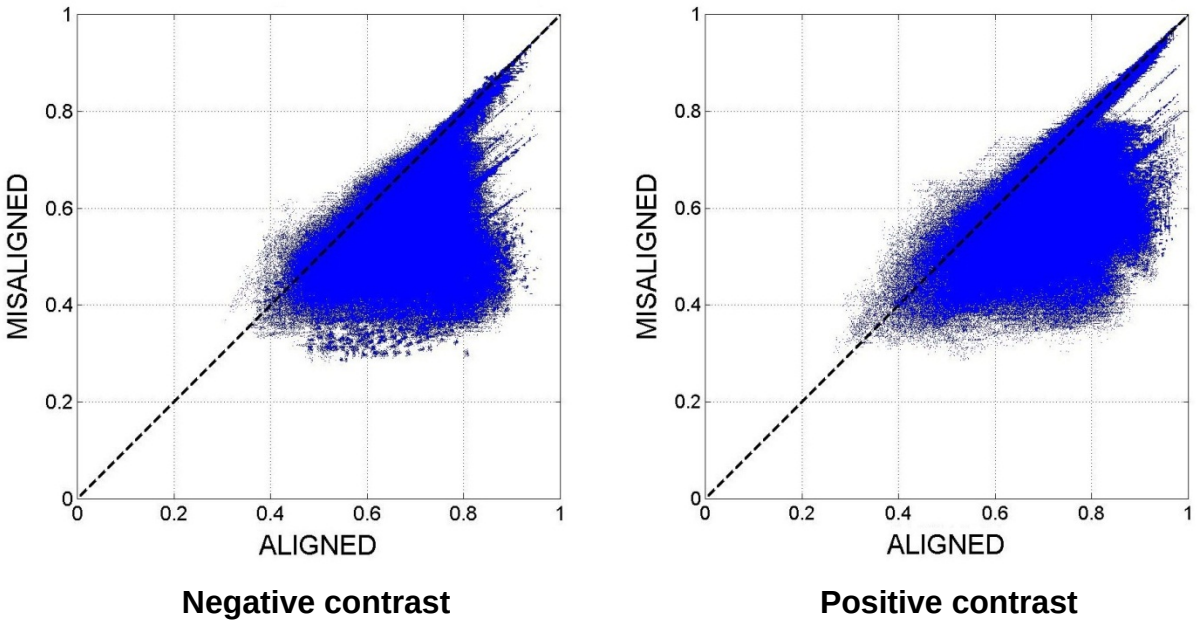


Figure 8.7. Scatter-plot of responses to misaligned versus aligned composites for negative (left) and positive (right) contrast.

8.3.2 Effect of contrast reversal on distances

Since contrast reversal did not strongly reduce individual unit responses, it stands to reason that the distances between negative faces are similar to distances between positive faces. Furthermore, since misalignment has similar effects on responses (to positive and negative faces), then it also stands to reason that misalignment will have similar effects on distances also. These two phenomena are indeed seen empirically in our model, and negative faces produce a misalignment effect for large, coarse features (Fig. 8.8), but not for small, fine features (Fig. 8.9)

8.4 Effects of contrast reversal on recognition

After accounting for the CFE, we now see if our model can also account for recognition performance. In this section, we look at model unit responses to the same set of images as before (i.e. composites with a gap between top and bottom halves). This is so that any differences cannot be attributed to the use of different stimuli. Note, however, that subjects in the recognition tasks as not instructed to ignore the bottom halves. Accordingly, we do not perform any attentional modulation on the images.

From Fig. 8.10, we see that on average, the distance between the positive and negative versions of the same face (Fig. 8.10 top row) is larger than even the distance between different faces that are compared within polarity (Fig. 8.10 third and fourth rows). This explains why performance in the PN and NP conditions are worse than the PP condition for faces (Finding #2).

What about objects? To examine this, we use small, fine features, for "object-like" processing. This time, however, the model does not match the empirical results (Finding #3). From Fig. 8.11, we see that, just as for "face-like" processing, contrast-reversed versions of the same face are even more different than different faces within a polarity. Does this mean our model is wrong? Not necessarily so.

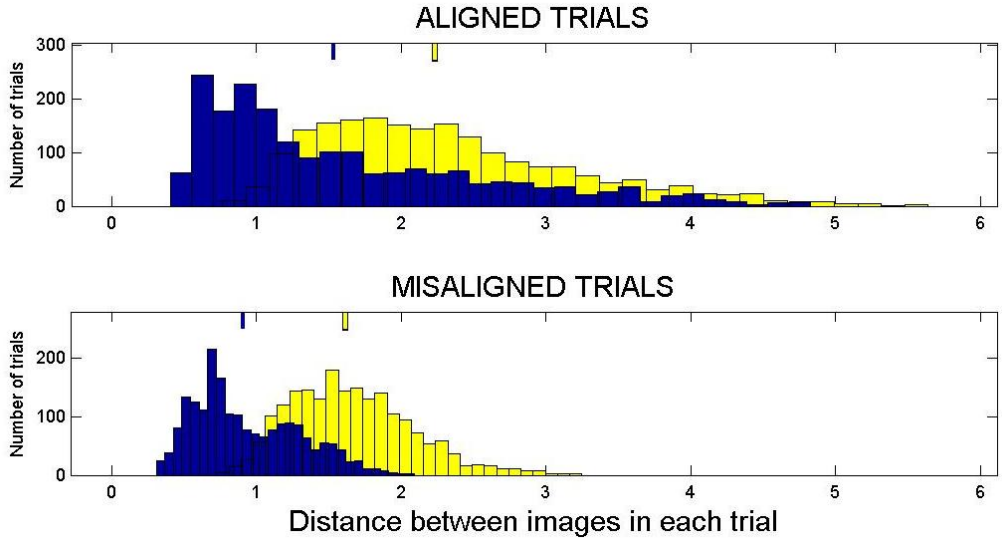


Figure 8.8. Histograms of distances for large, coarse features for contrast-reversed faces. Blue: "same" trials. Yellow: "different" trials. Hanging bar indicates mean.

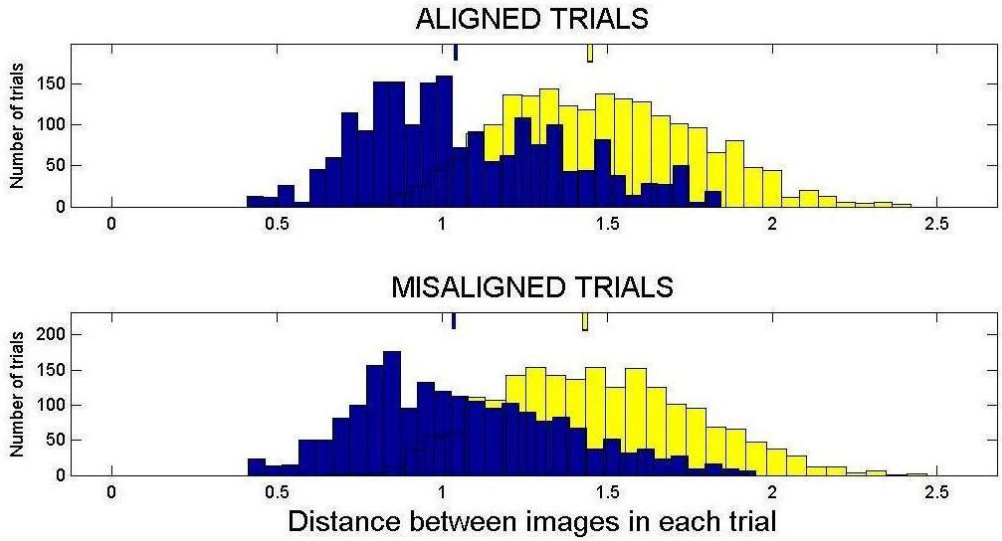


Figure 8.9. Histograms of distances for small, fine features for contrast-reversed faces. Blue: "same" trials. Yellow: "different" trials. Hanging bar indicates mean.

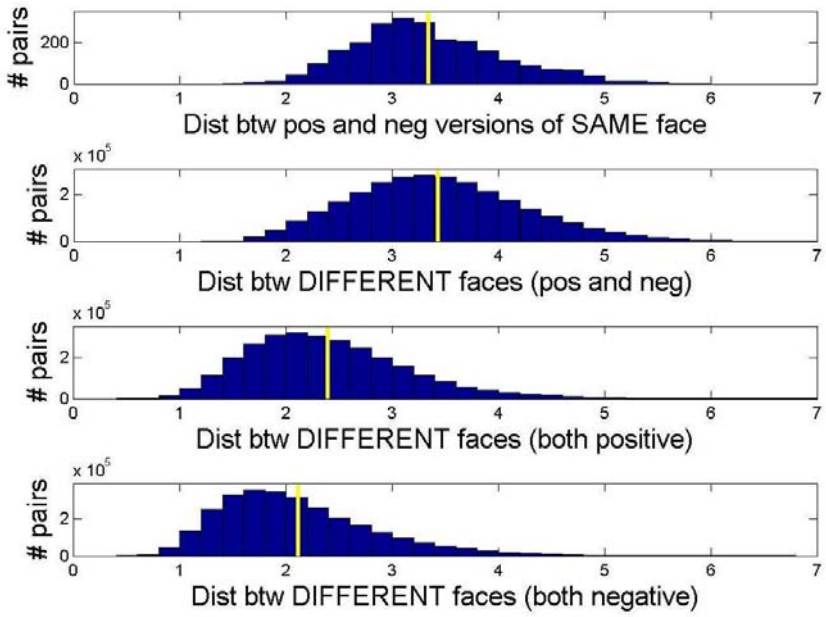


Figure 8.10. Histograms of distances between faces for various types of face pairs for large, coarse features. Yellow line indicates mean of distribution. Pos: positive. Neg: negative.

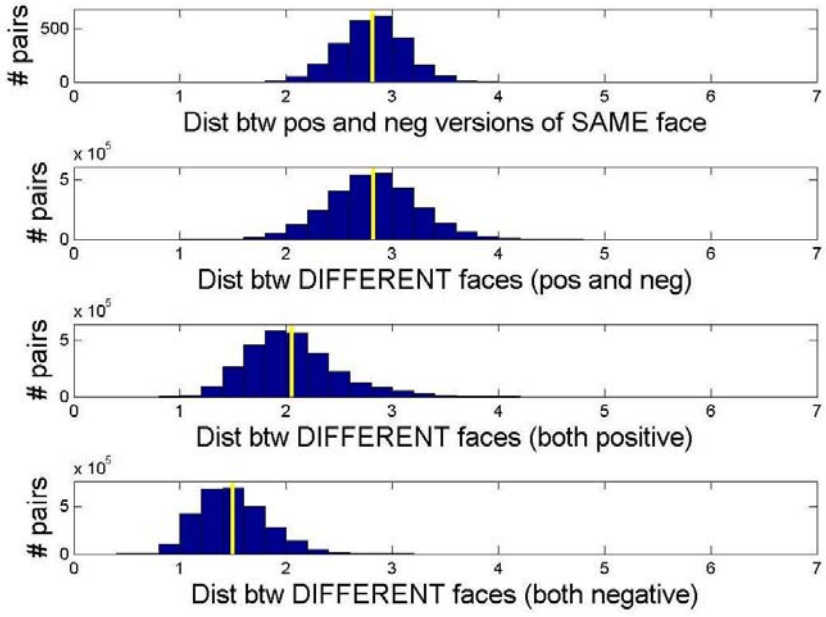


Figure 8.11. Histograms of distances between faces for various types of face pairs for small, fine features. Yellow line indicates mean of distribution. Pos: positive. Neg: negative.

First, from Fig. 8.12, we see that for inverted faces (using regular large, coarse features), the effect of contrast reversal is less than that for upright faces (compare top two rows versus bottom two rows for Figs. 8.10 and 8.12). This suggests that stimulus characteristics may play a larger role than type of processing in this context. Interestingly, Russell et al. (2006) found that for faces that differed only in shape but not pigmentation, faces of either contrast elicited equal performance. However, for faces that differed only in pigmentation but not shape, negative faces elicited significantly worse performance than positive faces. While this study compared PP versus NN conditions (rather than PN/NP versus PP, which we are examining here), it nonetheless supports our explanation. We therefore predict that especially for discrimination tasks that minimize memory effects, if exemplars within an object class are made to differ only in "pigmentation", then contrast reversal will cause a performance detriment, much like for faces. As a corollary, we claim that in studies that reported no effect of contrast reversal for objects, exemplars in these studies differed primarily (or at least diagnostically) in shape. In other words, the different findings for faces and objects arose from stimulus and task factors, rather than different processing mechanisms. Consistent with our prediction, Vuong et al. (2005) found that the addition of pigmentation cues lead to greater contrast reversal effects (performance difference between PP/NN and PN/NP) for both faces and "Greebles".

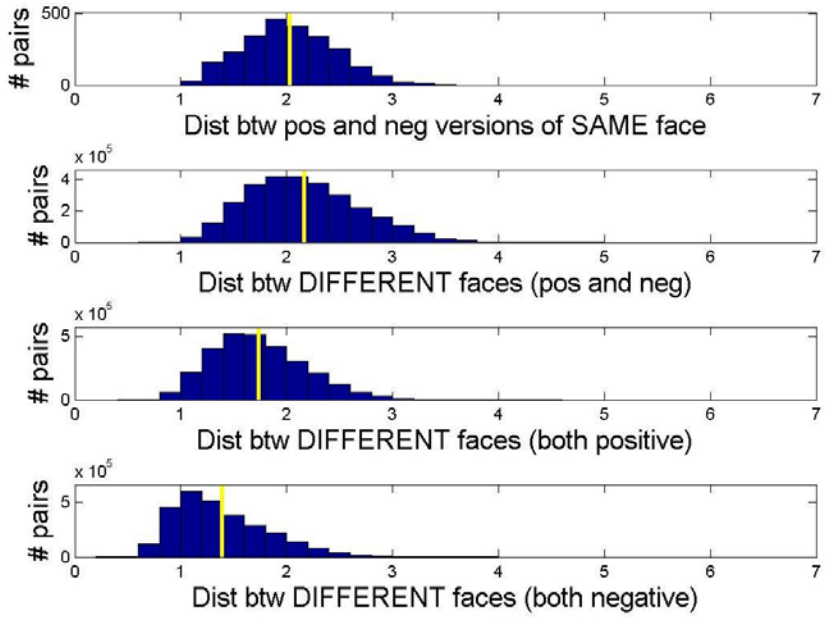


Figure 8.12. Histograms of distances between inverted faces for various types of face pairs for large, coarse features. Yellow line indicates mean of distribution. Pos: positive. Neg: negative.

We now turn to Finding #4 (comparison of NN versus PP). Robbins & McKone (2007) and Russell et al. (2006) found significant differences, while Liu & Chaudhuri (1997) did not. Can our model reconcile these results? There are many differences between the three studies, but we believe that the most salient difference is in the blocking (or not) of conditions. Conditions were

intermixed in the studies that found significantly worse performance for negative faces, while conditions were blocked in the study that did not. How is this pertinent to our model?

Note that in Fig.8.10, although the distance histograms for the PP and NN conditions (third and fourth rows, respectively) look fairly similar, if the same threshold is used in both conditions to determine if faces are same or different, then the PP condition does indeed lead to better performance than the NN condition (Fig. 8.13 left). Hence, intermixing of conditions (which limits different thresholds or "biases" for different conditions) is linked to better PP than NN performance. On the other hand, if conditions are blocked, then each condition may have its own threshold that optimizes performance for that condition, leading to smaller differences between conditions. In the case of Liu & Chaudhuri (1997), which had conditions blocked, there is in fact a small (but non-significant) performance drop in the NN condition compared to the PP condition. If conditions had been intermixed, this drop may have become significant, as predicted by our model, and as found in Robbins & McKone (2007) and Russell et al. (2006).

Furthermore, from Fig. 8.13 (left, red curve), we see that the performance difference between PP and NN conditions (a.k.a "reversal effect") tapers off at low and high accuracy levels. In Liu & Chaudhuri (1997), performance approached ceiling (91.4% for PP, 86.6% for NN). Thus, ceiling effects are another possible explanation for why they did not find a "reversal effect". In sum, the conflicting results are likely to be merely quantitative differences arising from either ceiling effects or procedural differences between studies.

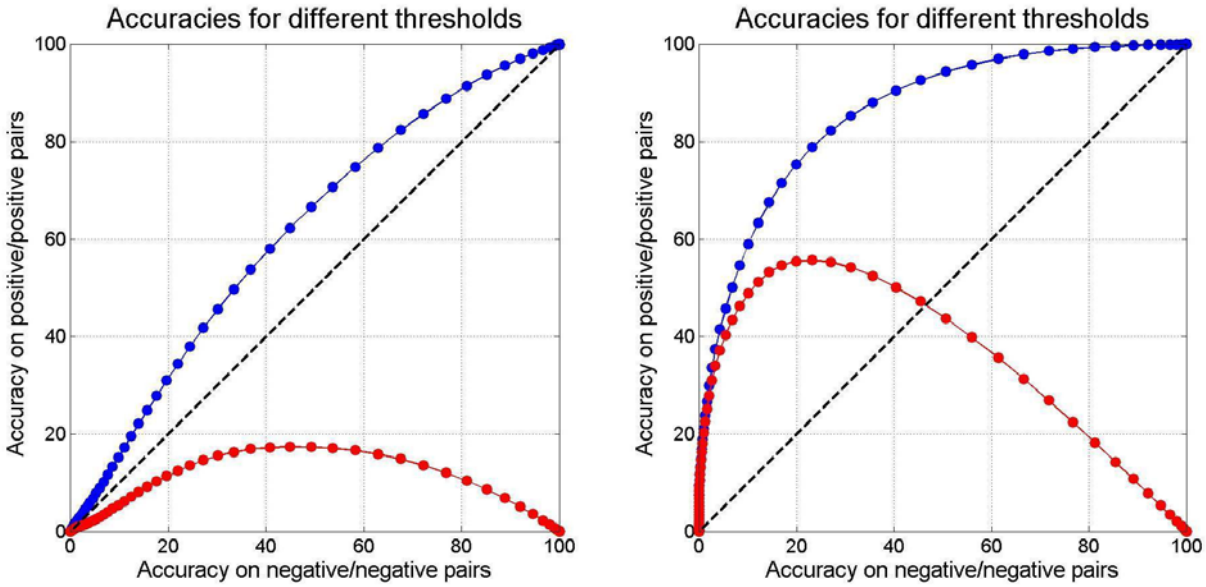


Figure 8.13. Accuracies for different thresholds. One dot for each threshold. Blue: accuracy for PP versus NN pairs. Red: magnitude of "reversal effect" (accuracy for PP minus accuracy for NN). Left: large, coarse features. Right: small, fine features. Dashed black line: equal accuracy for PP and NN.

Interestingly, Fig. 8.13 (right) shows that our model predicts that "object-like" processing also produces a "reversal effect" (possibly even larger; but this needs further investigation). Consistent with this, Robbins & McKone (2007, Fig. 10) found reversal effects for dog images in dog novices (but the effect size was smaller than for faces). We therefore predict that for faces, activity in areas such as LOC (and perhaps the OFA), reversal leads to poorer discrimination (measurable perhaps through adaptation).

8.5 Contrast reversal versus inversion

Interestingly, although our model predicts that inversion reduces the PN/NP versus PP/NN difference (compare the top two rows versus bottom two rows of Figs. 8.10 and 8.12), it also predicts a PP versus NN difference for inverted faces that is about as large (perhaps even larger) as for upright faces (Fig. 8.14). Robbins & McKone (2007) found precisely this; the PP versus NN "reversal effect" was highly significant ($p<0.001$) for both upright and inverted faces.

Robbins & McKone (2007, p.62) concluded from prior work that "these findings argue that contrast reversal effects and configural processing arise from different stages of visual processing". Our model replicates their results, but clearly shows that a single stage of visual processing can account for both types of effects.

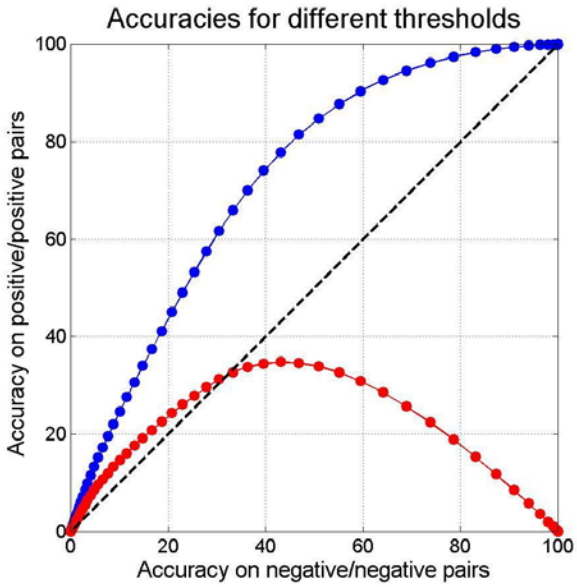


Figure. 8.14. Accuracies for inverted faces (large, coarse features). One dot for each threshold. Blue: accuracy for PP versus NN pairs. Red: magnitude of "reversal effect" (accuracy for PP minus accuracy for NN). Dashed black line: equal accuracy for PP and NN.

8.6 Chapter summary

In this chapter, we have used our model to both analyze and synthesize the body of results pertaining to contrast reversal. Consistent with empirical behavior, we have shown that although negative faces are less well discriminated than positive faces, they are nonetheless processed "holistically", as evidenced by the CFE.

Furthermore, we have demonstrated that the notion that "inversion disrupts processing" is not a useful one. Inverted faces elicit a significantly smaller "misalignment effect" than upright faces, but they elicit a "reversal" effect that is as strong as for upright faces. In all cases, identical processing occurs; it is simply the stimulus changes that give rise to these effects. Our model thus shows that a step-by-step, mechanistic understanding is crucial.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 9: Spatial Frequency

Chapter abstract

In this chapter, we look at the effect of spatial frequency filtering on face processing. Specifically, we attempt to reconcile the conflicting findings regarding the CFE for high spatial frequency (HSF) filtered faces. Overall, since our model uses coarse templates, we find that low spatial frequency (LSF) filtered faces are more "holistically processed" than HSF faces, consistent with other studies.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 9: Spatial Frequency

This chapter examines the issue of spatial frequency (SF), primarily focusing on the CFE for full spectrum (FS), low spatial frequency (LSF) filtered, and high spatial frequency (HSF) filtered faces. The issue of SF has been closely linked to face processing (see Ruiz-Soler & Beltran 2006 for a review). Interestingly, links between SF, face recognition and certain disorders like Autism have been found (e.g. Deruelle et al. 2004, Leonard et al. 2011).

In Chapter 5, we showed that largeness was the key factor in producing the CFE, rather than spatial scale. We nonetheless maintained our hypothesis that the characteristics of face processing arise due to the use of large, coarse templates (and small, fine templates for object-like processing). This was for several reasons, including the maintenance of constant "complexity" (number of afferent C1 units) as a control, and also in line with the idea of "informativeness" (Ullman et al. 2002); see Section 13.4.6 for a discussion of more reasons.

In the previous chapter, we indirectly alluded to the coarseness (i.e. SF) of the templates as a factor that enabled our model to account for the characteristics of face processing for contrast-reversed faces.

In this chapter, we directly examine this issue, especially in relation to the CFE. Since our model uses coarse templates, one might expect that differently filtered images may be processed quite differently by our model. However, like with many other issues relating to face processing, the issue of SF filtering for the CFE has yielded conflicting behavioral results. We show that our model can reconcile these results.

9.1 The CFE and spatial frequency

Several studies have investigated the effect of SF filtering on the CFE (Goffaux & Rossion 2006, Cheung et al. 2008, Goffaux 2009), but have found conflicting results. All three studies used 8 cycles per face (cpf) as the low frequency cut-off, and 32 cpf as the high frequency cut-off. Fig. 9.1 shows the FS, LSF and HSF versions of one example face. Note that oval-cropping and insertion of the mid-line gap were done post-filtering, to maintain the sharpness of the face boundary and midline gap, as was done by Goffaux & Rossion (2006) and Cheung et al. (2008).

(Methodological note: we attempted to replicate the SF filtering methods of the aforementioned studies as faithfully as possible, but some essential details were omitted. Furthermore, as is apparent from the figures in the three studies, their stimuli were rather different)

Goffaux & Rossion (2006) found that LSF faces had a larger misalignment effect than HSF faces. Cheung et al. (2008) replicated these results, but showed that when the "complete" design is used, there is no difference in the congruency effect (and congruency x alignment interaction) for LSF and HSF faces. However, Goffaux (2009) also used the "complete" design, but showed that there is a significant difference between LSF and HSF faces in terms of congruency effect and (congruency x inversion) interaction.

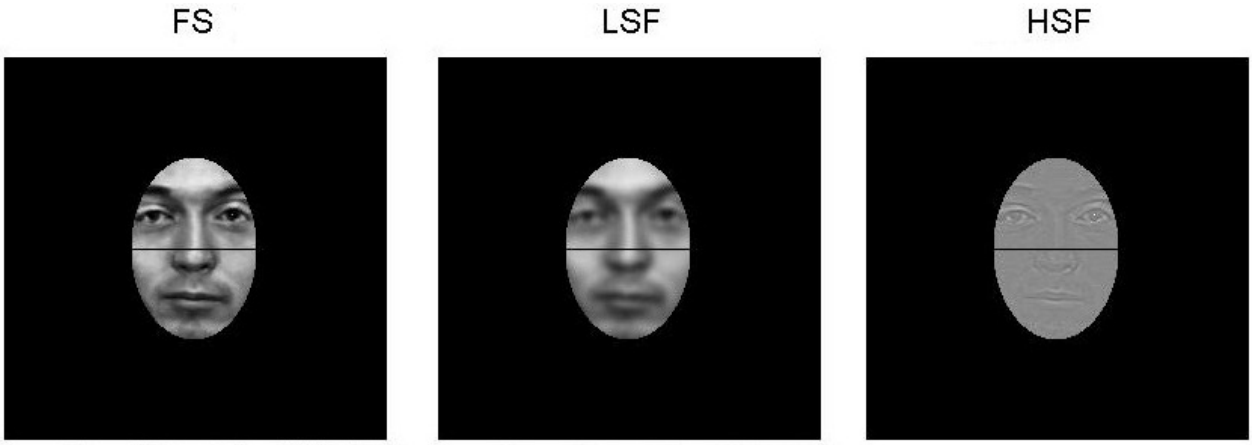


Figure 9.1. Full-spectrum (FS), low spatial frequency (LSF), and high spatial frequency (HSF) versions of a typical composite face.

Why did these studies find opposite results? Cheung et al. (2008) replicated the methods of Goffaux & Rossion (2006) quite faithfully, so it is not surprising that they also replicated their results (for the "partial" design misalignment effect). (But note that their filtered stimuli appear to be visually somewhat different, at least from the published images)

Goffaux (2009) used very different experimental conditions. In particular, they used a simultaneous presentation of both composites, and presentation time was up to 3 seconds (compared to 600ms/1000ms sequential presentation for the other two studies). In addition, the stimuli of Goffaux 2009 were much larger (6° x 7.8° versus 3.1° x 4.1°). Moreover, the filtered faces were adjusted to match the unfiltered FS faces in luminance and RMS contrast, something that the other studies did not do.

With so many differences between Goffaux (2009) and the other two studies, is it hopeless to try to reconcile them? Should we just conclude that the relationship between SF and face processing is not a robust one? In the next section, we show that one very simple factor may be sufficient to reconcile all three sets of results.

9.2 Reconciling the conflicting studies

We first begin by using the "complete" design, and attempt to replicate the results of Cheung et al. (2008). Fig. 9.2 shows the D' for FS, LSF and HSF faces for various thresholds. As discussed in Chapters 5 and 7, the threshold affects only the magnitude, not existence, of the CFE (though the CFE obviously disappears for extreme thresholds). The reason why we display many thresholds here will be apparent later.

When comparing FS, LSF and HSF results, we are making judgments about the magnitude of the CFE, not its existence (all studies found the CFE, even for HSF faces). We must therefore be

careful about choosing which threshold to use. Both Goffaux and Rossion (2006) and Cheung et al. (2008) randomly intermixed trials types. This would suggest that subjects apply the same threshold from trial to trial, and it should not differ (much) for FS, LSF and HSF faces. But which threshold to use? For now, we make the assumption that subjects implicitly choose a threshold that maximizes overall performance. From the average D' (yellow lines) in Fig. 9.2, we see that the best thresholds are roughly 0.750, 0.450 and 0.900 for LSF, HSF and FS faces respectively. Therefore, the optimal threshold over all SF conditions would be roughly between 0.600 and 0.750. We thus focus on these two thresholds in Fig. 9.3.

From Fig. 9.3, we see that the magnitude of the CFE is similar, regardless of SF condition (and CFE metric), qualitatively replicating the results of Cheung et al. (2008). This is particularly true for the 0.600 threshold. For the 0.750 threshold, the CFE for HSF faces is slightly less, but in practice, noisy computation, experimental noise and inter-subject variation would most likely ablate these differences. More importantly, the CFE magnitudes are in the same ballpark, and are generally larger than the differences in magnitude (this is more so for the congruency effect than the congruency x alignment interaction).

On the other hand, Goffaux (2009) found that HSF faces elicited a significantly smaller CFE. As discussed earlier, there were many differences between the studies (Cheung et al. 2008 and Goffaux 2009), but we would nonetheless like to see if our model can reconcile these results. As hinted at in the previous paragraph, we believe that the threshold is the key factor.

Figure 9.4 shows the (congruency x alignment) effect for LSF and HSF faces at a broad range of thresholds. (The congruency effect at different thresholds is already shown in Fig. 9.2). It is now very clear that the roughly equal CFE magnitude for LSF and HSF faces is only true for a very small range of thresholds (roughly around 0.600). For LSF faces, the CFE magnitude keeps increasing up to even a threshold of 1.650, whereas for HSF faces, the CFE magnitude peaks at 0.600. Thus, even for a moderately different threshold from before, like 0.900, CFE magnitude (for upright faces) for LSF faces becomes more than twice that for HSF faces. In short, the CFE is equal in magnitude for LSF and HSF faces only in some circumstances.

Note that we are not suggesting that a different threshold is literally the only difference between the two studies, nor that Cheung et al. (2008) found a spurious result. It is not at all clear what methods are actually used in the human brain to determine thresholds. The usefulness of models (such as ours) in situations like this, is to allow a simulation and examination of the effects of various parameters to understand the range of possible outcomes. What we have found is that even with a system using coarse templates, there exist reasonable conditions (e.g. maximal mean D') in which both LSF and HSF faces might seem to be roughly equally holistic. Under a broader range of conditions, however, LSF faces generally elicit larger CFEs than HSF faces, which is not surprising for our model.

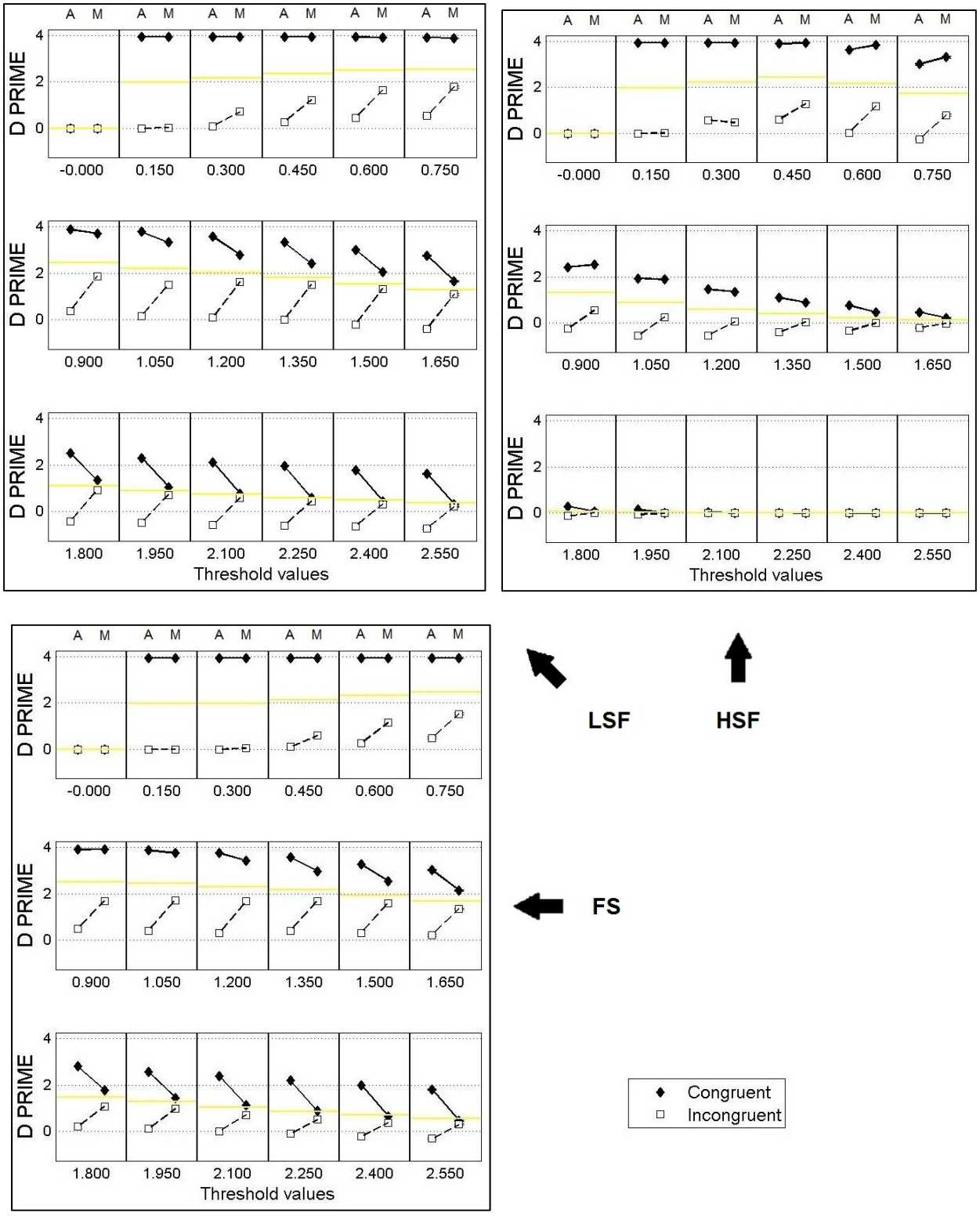


Figure 9.2. CFE for large, coarse templates over various thresholds (numeric values indicated on x-axis). Top left: LSF faces. Top right: HSF faces. Bottom: FS faces. A: aligned. M: misaligned. Yellow line: mean D' over all four conditions for each threshold.

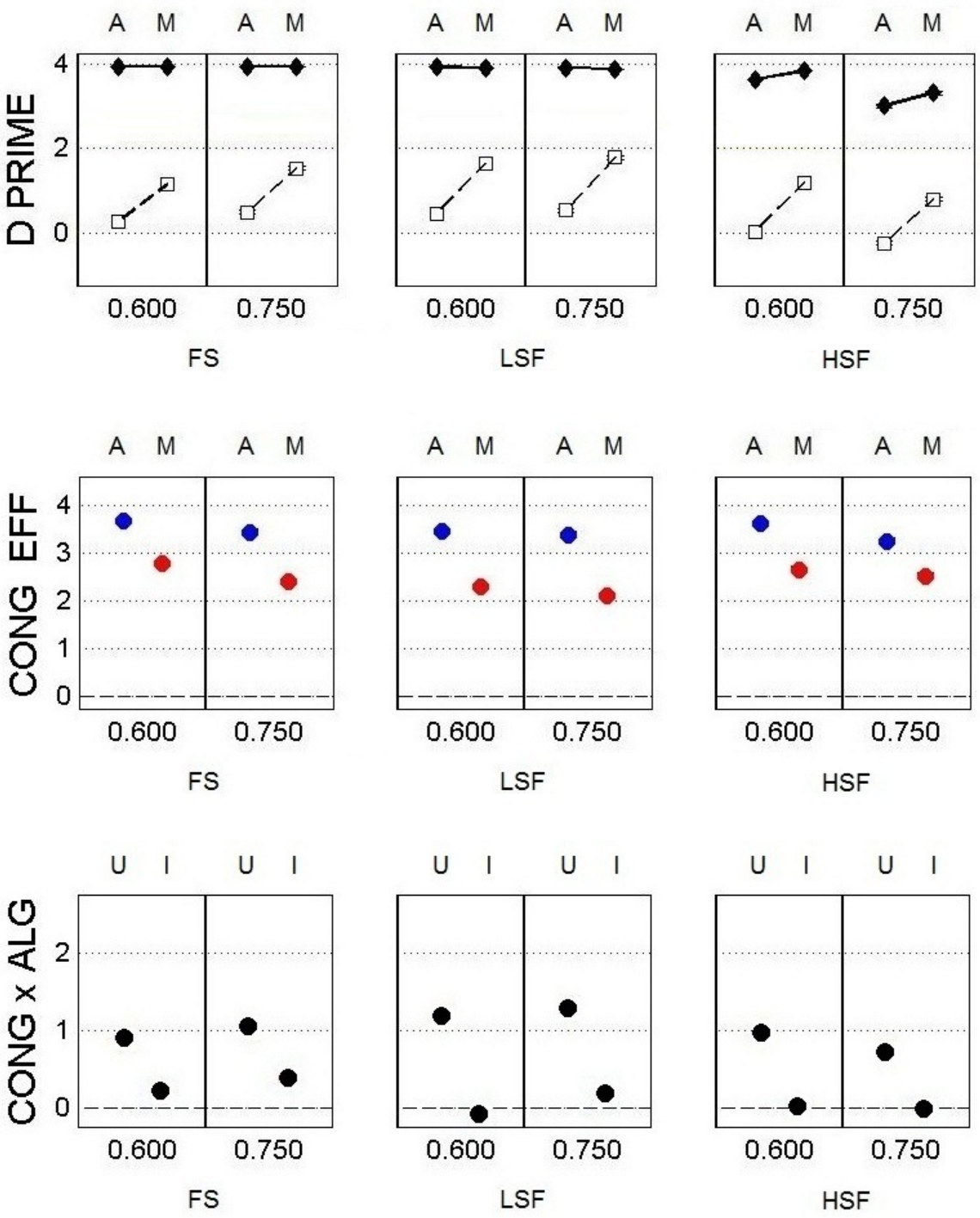


Figure 9.3. The CFE for FS (left column), LSF (middle column) and HSF (right column) faces at two thresholds (0.600 and 0.750), measured by three metrics. Top row: D' (black diamond: congruent, white square: incongruent). Middle row: congruency effect (i.e. congruent D' – incongruent D'). Bottom row: congruency x alignment interaction (i.e. aligned congruency effect – misaligned congruency effect). A: aligned. M: misaligned. U: upright. I: inverted.

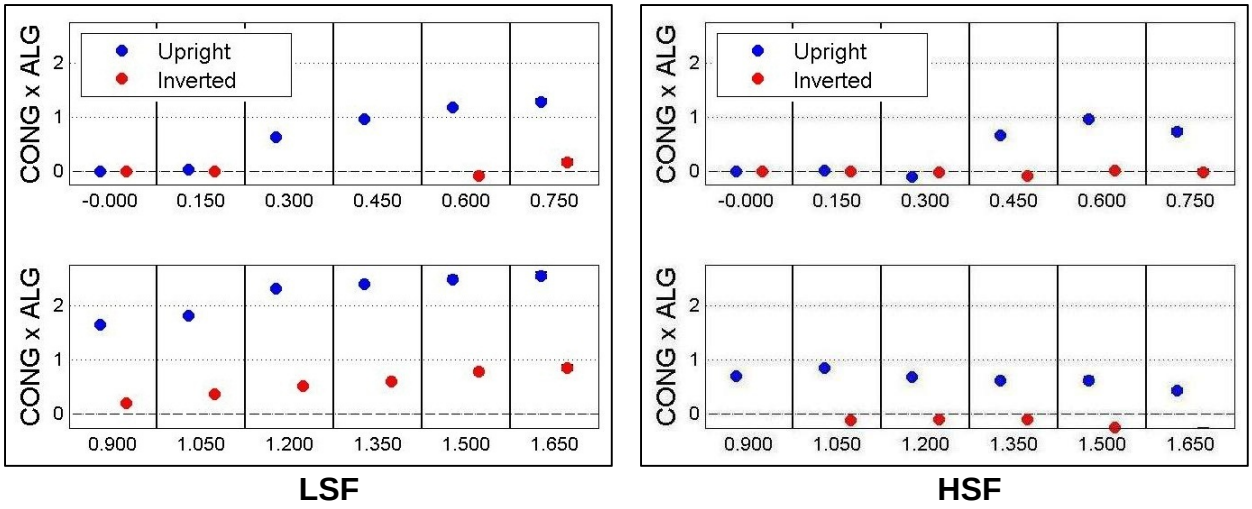


Figure 9.4. CFE as measured by (congruency x alignment) interaction for a broader range of thresholds (numeric values indicated on x-axis). Left: LSF faces. Right: HSF faces. Blue: upright faces. Red: inverted faces.

Before we turn to a more in depth exploration of how and why these SF differences come about (in the next section), we note an interesting "post-diction" made by our model that strikingly matches the behavioral data qualitatively. Cheung et al. (2008) claimed that the "partial" design used by Goffaux & Rossion (2006) is susceptible to some poorly-understood "biases" that the "complete" design is able to sidestep through the use of D', and that it is these biases that gave rise to the supposedly incorrect finding of greater CFE for LSF than HSF faces.

Using the same thresholds as before (0.600 and 0.750), we found that our model produces the same qualitative behavior that Cheung et al. (2008) find in their calculation of bias (Fig. 9.5). Specifically, they found that congruent trials have a more negative bias than incongruent trials (marginal significance of $p=0.074$), misalignment shifts biases in the negative direction (main effect of alignment, $p=0.0001$), and biases are more towards the negative direction for HSF than LSF faces (main effect of SF, $p<0.0001$). Importantly, there is an interaction between SF and alignment ($p<0.01$).

As we will see in the next section, these differential biases can be explained rather simply as arising from the distances between faces, not some arbitrary or unknown effects. Briefly, assuming similar thresholds across all conditions, the smaller distances for HSF faces, misaligned trials, and congruent trials lead to the more negative biases in all these cases.

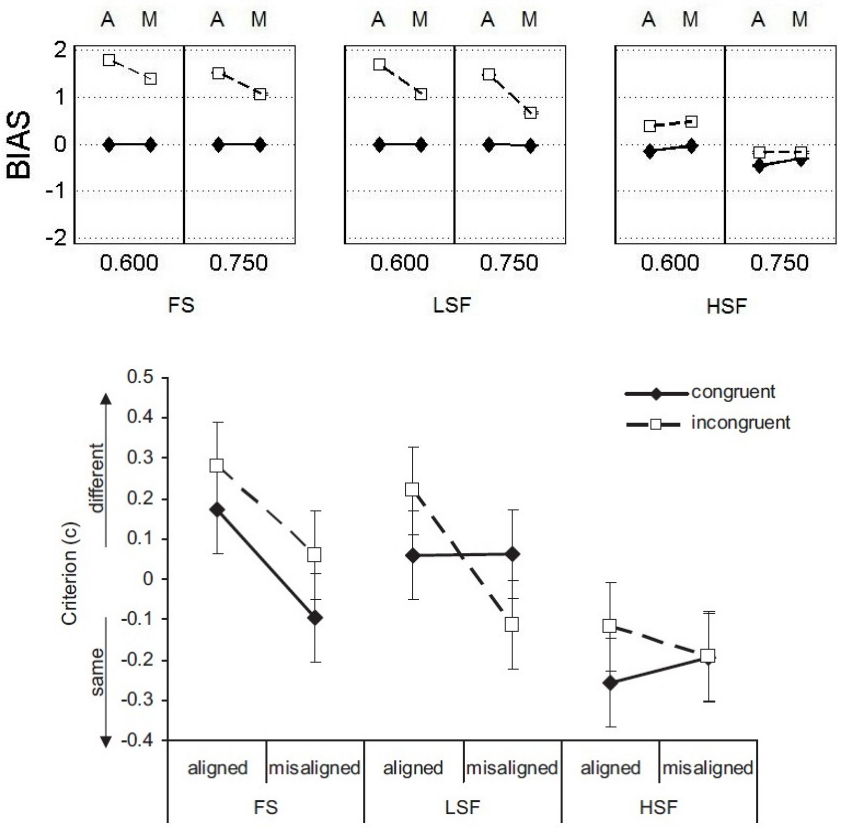


Figure 9.5. Bias (a.k.a. "criterion") for FS, LSF and HSF faces. Top row: model results for the two thresholds that correspond to the best overall average D'. Bottom row: behavioral results (reproduced from Cheung et al. 2008. See p.16 for copyright notice).

9.3 Step-by-step account

After looking at the CFE results, we now examine in more detail how these results come about. As always, we start from the responses of individual model units, and then look at the distance between two faces (calculated using the responses over a population of units).

9.3.1 C1 responses

We first look at the "physical properties" of these SF-filtered images, in terms of the V1-like (C1) responses that they elicit. From Fig. 9.6, we see that at the coarse C1 scales (e.g. scale 7), LSF responses are more similar than HSF responses are to the FS responses. At the finer scales (e.g. scale 1), the converse is true. (Note that the differences are visually subtle, partly due to averaging over orientations; the subtlety may be worsened by display factors when viewing on different computers or printing to different printers)

However, it is also important to note that the highest responses are mostly at the face boundaries, which are very similar for all SF conditions. In other words, at least for the stimuli that we use

(also used in Goffaux & Rossion 2006 and Cheung et al. 2008 – and possibly other studies too), the differences between SF conditions could be reduced due to the presence of strong, sharp edges. (Importantly, such edges contain information at all frequencies, not just the HSFs, and therefore affect all scales strongly)

Given that the C1 responses at coarse scales for LSF and FS are more similar than HSF and FS are, one would then expect that the same would be true for C2 responses corresponding to coarse templates (templates which are simply snapshots of C1 responses at coarse scales). We examine these responses in the next section.

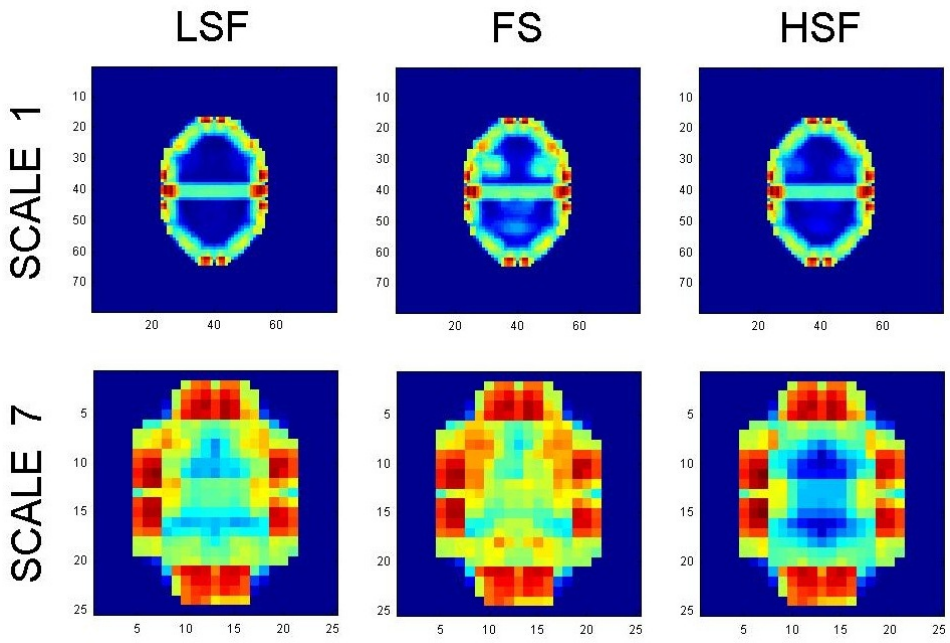


Figure 9.6. C1 responses (averaged over all orientations) at scales 1 and 7 to LSF, FS and HSF versions of a composite. Blue: low activity. Red: high activity. Note: the differences between LSF and HSF responses were less noticeable for scale 3, so scale 1 is shown instead (this is purely due to visualization reasons).

9.3.2 C2 responses

Figure 9.7 (top left) shows that as expected, the mean responses to FS faces are highest, followed by LSF, and then HSF faces. It should also be noted, however, that the differences are moderate, not large. The scatter-plots in Fig. 9.7 confirm these findings: the responses to all SF conditions are highly correlated, but overall the largest differences are between FS and HSF responses.

For small, fine features (results not shown), as expected, the opposite results are found. The largest differences are between FS and LSF responses, and the mean responses to LSF faces are the lowest.

Apart from the responses in the different SF conditions per se, what about the effects of misalignment? Fig. 9.8 shows that the effect of misalignment (i.e. difference in response to aligned vs. misaligned counterparts) is larger for LSF than for HSF faces. Again, the difference between LSF and HSF faces is moderate (but noticeable).

The differences at the level of individual model units are moderate, but we will see in the next section that when using many units to calculate distances between faces, the differences can be quite large.

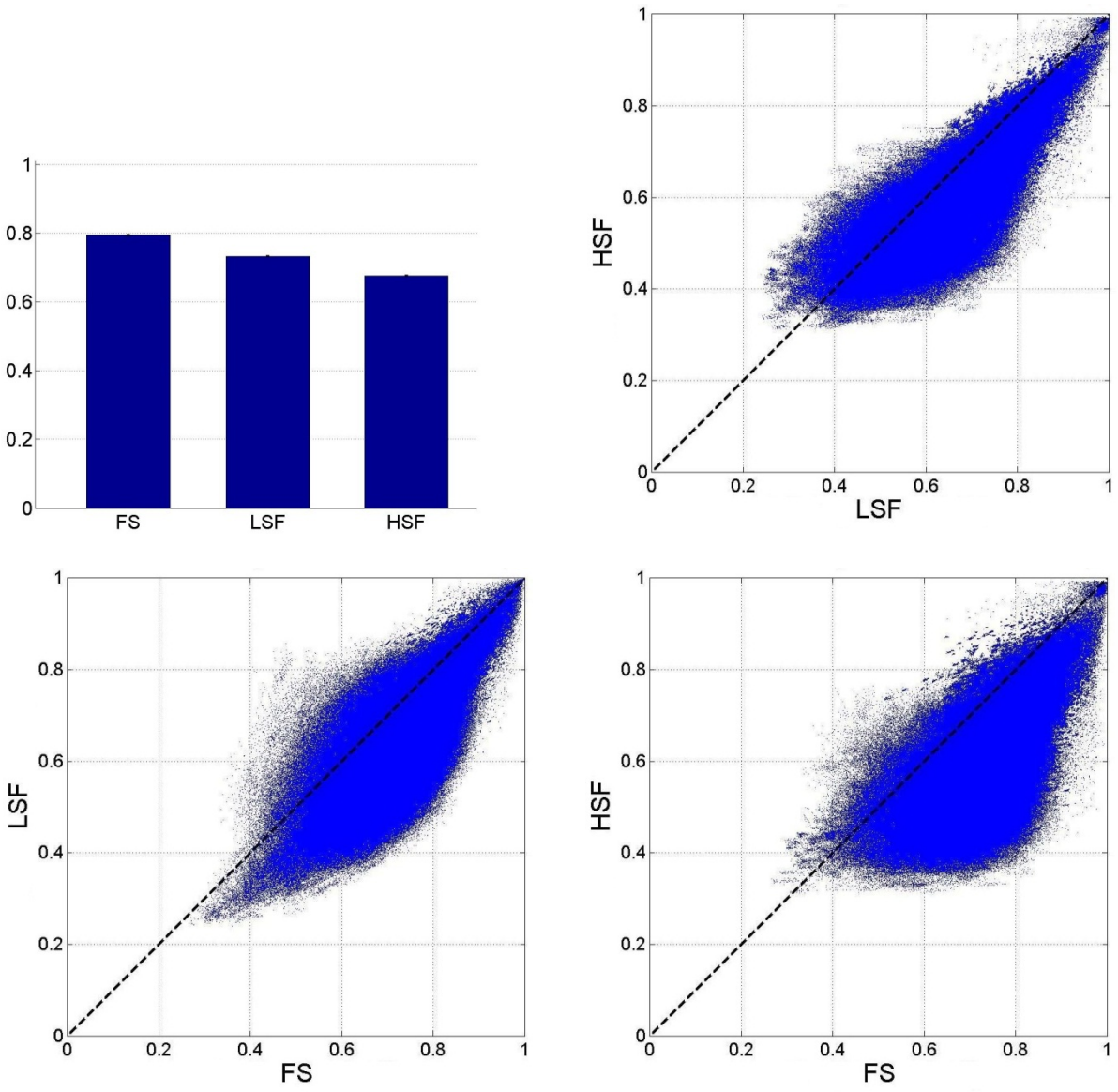


Figure 9.7. Responses of large, coarse templates to FS, LSF and HSF composites. Top left: mean responses (average of 1000 units x 2450 composites). Clockwise (from top right): scatter-plots of responses to HSF vs. LSF, HSF vs. FS and LSF vs. FS composites. Each scatter plot displays 1000 x 2450 points.

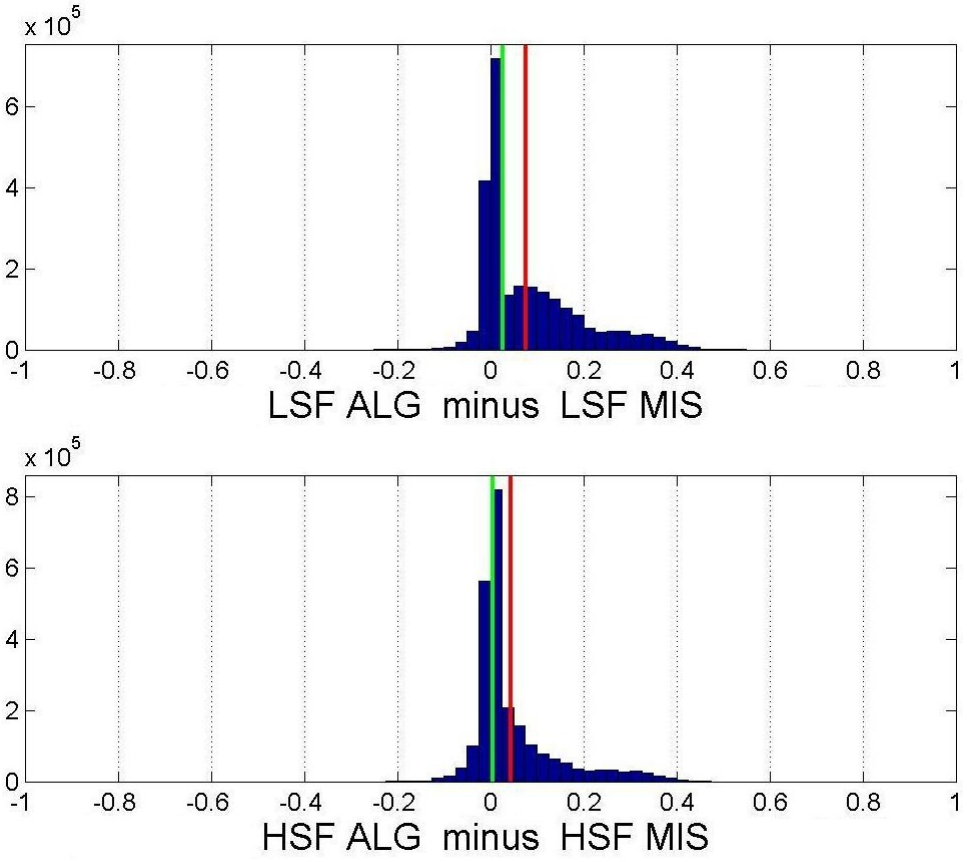


Figure 9.8. Histograms of effects of misalignment on responses. X-axis: response to aligned (ALG) composite minus response to misaligned (MIS) counterpart. Y-axis: number of units. Top: LSF. Bottom: HSF. Red line indicates mean of distribution. Green line indicates median.

9.3.3 Distances between images

Figure 9.9 shows the distribution of distances for the "partial" design. First, the average distance for LSF faces is roughly twice that for HSF faces (note the different x-axis scales). More importantly, however, is the effect of misalignment. The moderate difference in effect of misalignment for individual units (Fig. 9.8) is now much more apparent.

This difference is robust to specific threshold, but for illustrative purposes we have used a threshold of 0.7 (red lines in Fig. 9.9). For LSF faces (top panel), misalignment causes the proportion of "same" trials (blue) falling below the threshold (i.e. the hit-rate) to become much larger. This increase in hit-rate is much smaller for HSF faces (bottom panel), replicating Goffaux & Rossion (2006) and Cheung et al. (2008).

Interestingly, our model "predicts" that the hit-rate for HSF faces is generally higher than for LSF faces (if the same threshold is applied to both conditions), which was found empirically by both Goffaux & Rossion (2006) and Cheung et al. (2008).

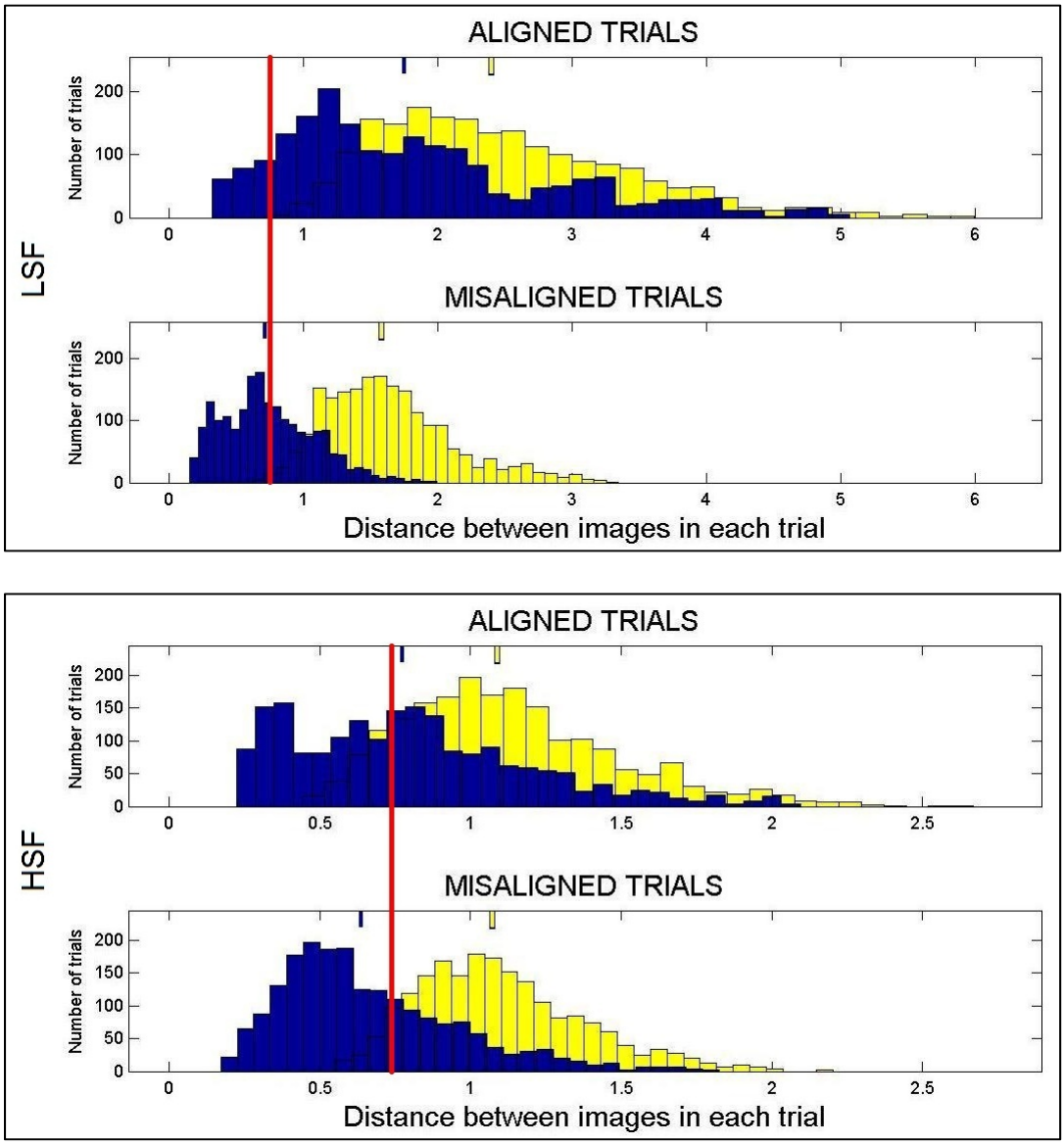


Figure 9.9. Histograms of distances between the two composites in each trial, for the "partial" design. Top: LSF. Bottom: HSF. X-axis: Euclidean distance (note different scales for LSF and HSF). Y-axis: number of trials. Blue: "same" trials. Yellow: "different" trials. Red line: arbitrary threshold, set to ~0.7 in this figure.

We now turn to the "complete" design, shown in Fig. 9.10. Since our model is noiseless, the congruent-same trials (which show identical faces) always have a distance of 0, and these are not shown (except by the hanging blue bar at distance 0). In the "complete" design, the main effect is the "congruency effect", i.e. the difference in D' for congruent trials (brighter shades) versus incongruent trials (darker shades).

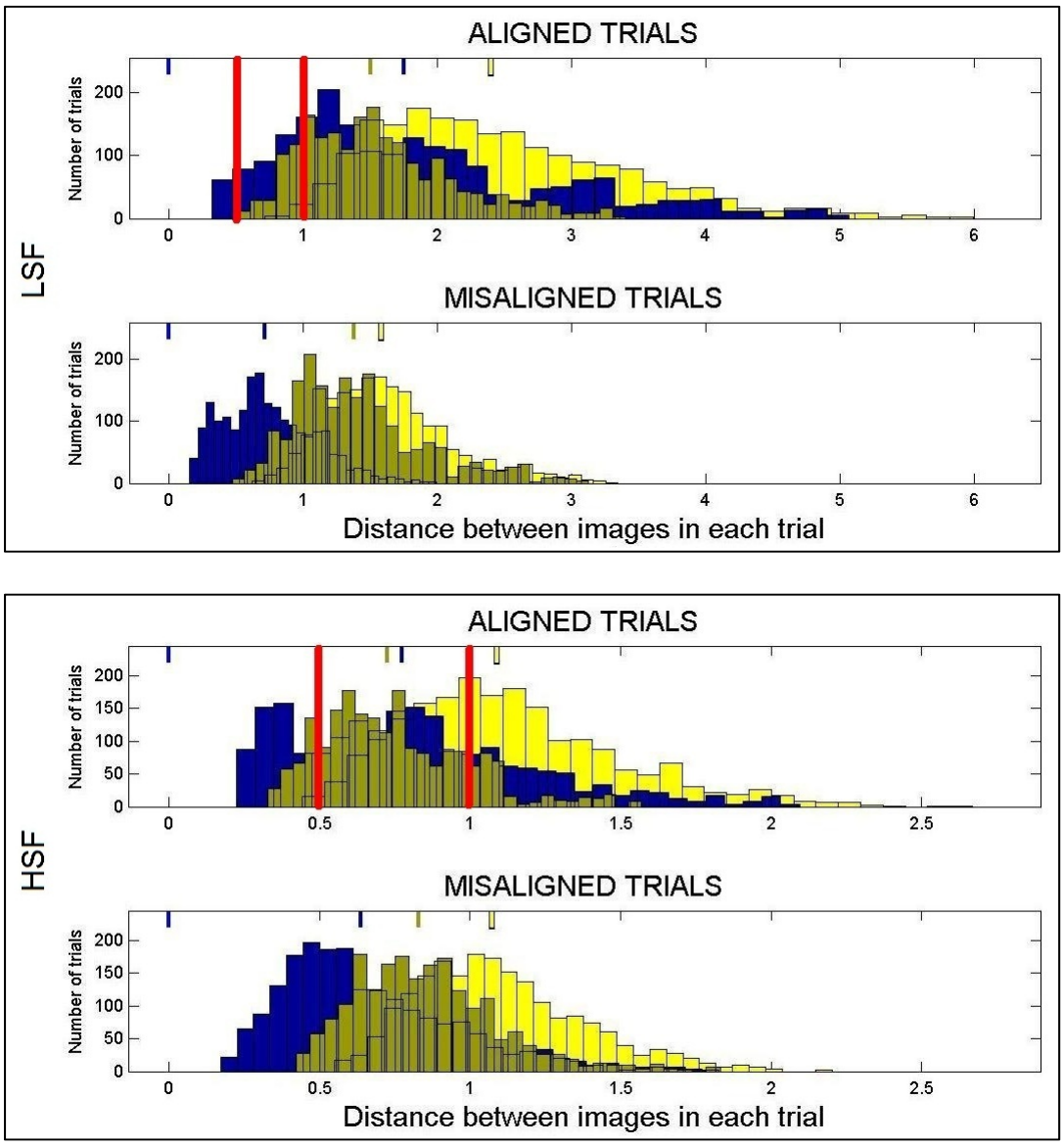


Figure 9.10. Histograms of distances between the two composites in each trial, for the "complete" design. Top: LSF. Bottom: HSF. X-axis: Euclidean distance (note different scales for LSF and HSF). Y-axis: number of trials. Blue hues: "same" trials. Yellow hues: "different" trials. Brighter shades: congruent. Darker shades: incongruent. Red line: arbitrary threshold, set to 0.5 and 1.0 in this figure.

For simplicity, we only discuss the aligned trials here. As discussed earlier, the threshold matters in the comparison of LSF and HSF conditions. For a low threshold like 0.5, we see that for both LSF and HSF, the vast majority of the congruent-different trials (bright yellow) are above the threshold, leading to very high congruent D' (see Fig. 9.2). For the incongruent trials (dark blue and dark yellow), their distribution of distances are similar, so the congruent D' is very low (and this does not change much regardless of threshold).

For a higher threshold like 1.0, we see that now there is a big difference between LSF and HSF. For LSF, because the distances are generally much larger, the threshold of 1.0 gives results that are not dramatically different from the 0.5 threshold. However, for HSF, the 1.0 threshold makes a big difference. Many of the congruent-different trials (bright yellow) are below the threshold, leading to a strong false-alarm rate and much lower congruent D' than before. From Fig. 9.2, we see precisely this: for a threshold of 1.050, the difference between LSF and HSF is primarily in the congruent D' (black diamonds).

9.4 Spatial frequency and object-like processing

Finally, as the usual control, we contrast the large, coarse, templates with small, fine templates. As mentioned earlier, the opposite results are found for small, fine templates. The C1 responses for HSF faces are more similar than LSF faces to FS faces, and the C2 responses to HSF faces are larger than to LSF faces (results not shown).

What about the CFE? We have already shown that for regular FS faces, small, fine templates do not produce a CFE (at least when defined by a congruency x alignment interaction). Nonetheless, we note that Fig. 9.11 shows that as one would expect, the congruency effect (not interaction) is smaller for LSF faces than for HSF and FS faces. In other words, for "object-like" processing, our model predicts that LSF images will show a smaller congruency effect (but not a smaller congruency x alignment interaction) than for HSF images.

9.5 Chapter summary

In this chapter, we examined the effects of spatial frequency filtering on the CFE. As would be expected for a system using coarse templates, the CFE is generally larger for LSF than HSF faces. We nonetheless show that under some (reasonable) assumptions, the CFE for LSF and HSF faces can be similar, thus reconciling the results of two conflicting studies.

More broadly, the key contribution of this chapter is in reinforcing the notion that both "partial" and "complete" designs can be accounted for, reconciled and equally valid. The issue of "biases" that cast doubt on the "partial" design can be accounted for by our model.

More general investigation of SF and holism (e.g. Goffaux et al. 2005, McKone 2009b) or general face (e.g. Hayes 1988, Vuilleumier et al. 2003) or object (e.g. Oliva & Schyns 1997) processing is left for future work.

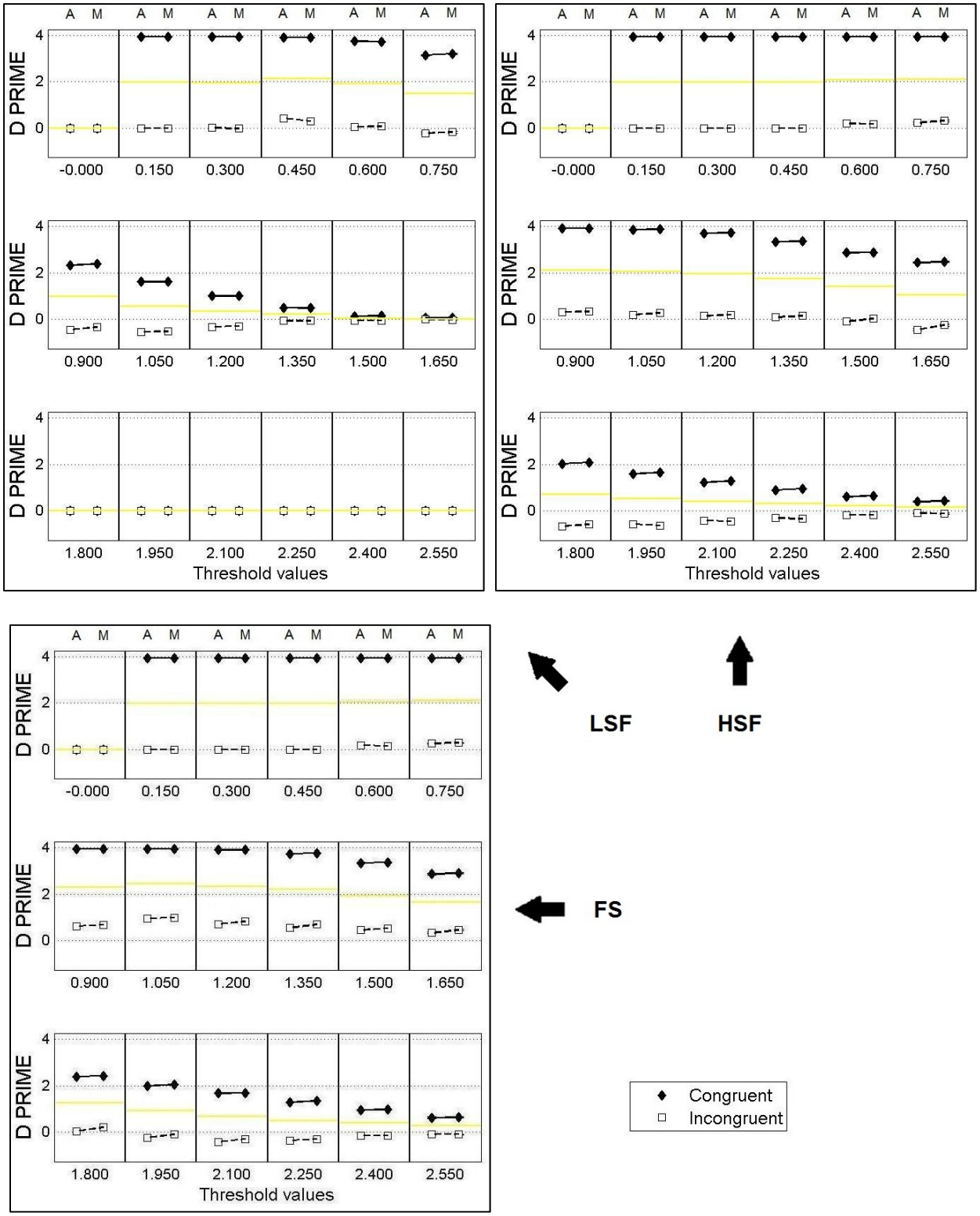


Figure 9.11. CFE for small, fine templates over various thresholds (numeric values indicated on x-axis). Top left: LSF faces. Top right: HSF faces. Bottom: FS faces. A: aligned. M: misaligned. Yellow line: mean D' over all four conditions for each threshold.

Chapter 10: The Face Inversion Effect (FIE)

Chapter abstract

The main purpose of this short chapter is to replicate the Face Inversion Effect (FIE), which is not the same as the CFE for inverted faces (Chapter 6). The key contribution of this chapter is in demonstrating the mechanistic relationship between holism and inversion, which surprisingly has not been shown to date.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 10: The Face Inversion Effect (FIE)

The main purpose of this chapter is to show that our model can also replicate the differential FIE. Note that our model is solely designed to be "holistic", and it is not clear from prior work what the exact relationship between holism and the FIE is. Our work will shed light on this relationship, by showing that large, coarse features ("face-like" processing) produce a larger FIE than small, fine features ("object-like" processing).

10.1 Model results

Here, we show that the model replicates the differential FIE, i.e. a larger inversion effect for "face-like" processing than "object-like" processing.

As before, we do not make any changes to the model used to demonstrate the CFE. Furthermore, we use the same composite images (but with no attentional modulation). We selected 49 composites such that none of the halves appeared in more than one composite, i.e. 49 distinct top halves were paired with 49 distinct bottom halves. For "face-like" processing, we use 1000 large, coarse features. For "object-like" processing, we use 1000 small, fine features.

We simulate a same-different discrimination task, whereby given a pair of images, the model has to determine if the images are same or different. This is similar to the tasks used by other studies, which reduce memory-related confounds. In our case, since the distance between images in the "same" trials will always be 0 (by definition, and due to noiseless conditions), we simply look at the "different" trials.

10.1.1 Step-by-step account: responses

As with the CFE, we give a step-by-step account of the differential FIE, beginning with the responses of individual model units. From Fig. 10.1, we see that as a result of inversion, the large, coarse units suffer a much larger drop in response than the small, fine units. For the large, coarse units, the mean response drops from 0.82 to 0.46. For the small, fine units, the mean response drops from 0.84 to 0.70. Note that the mean response to the upright faces is similar to both types of units (0.82 versus 0.84).

Interestingly, the correlation between upright and inverted responses is -0.44 for the large, coarse units. From Fig. 10.1, we see that the units with larger upright responses tend to suffer larger decreases. Consistent with this, the correlation between upright response and magnitude of decrease is 0.82 (results not shown). We are not aware of any studies investigating this relationship, so this is a novel prediction by our model.

10.1.2 Step-by-step account: distances

Next, we calculated the Euclidean distance between all pairs of faces. As Fig. 10.2 shows, inversion causes a larger decrease in distance for large, coarse features than small, fine features. For large, coarse features, the mean distance decreases from 2.62 to 1.82 (a drop of 0.80). For small, fine features, the mean distance decreases from 2.06 to 1.76 (a drop of 0.30). Instead of

looking at the change in mean distance, we also looked at the mean change in distance (results not shown). The numbers were essentially identical (mean decrease of 0.80 for large, coarse features and 0.30 for small, fine features).

Since we are only looking at the "different" trials (distances between pairs of composites that are different from one another), a drop in distance implies a decrease in discriminability. In other words, the findings above imply that inversion causes a greater decrease in discriminability for large, coarse features than for small, fine features. This is essentially the differential FIE.

10.1.3 Step-by-step account: accuracies

We go on to actually calculate accuracies and gauge the size of the FIE. To get accuracy values, we assume a fixed threshold that determines whether a pair of images is considered "same" or "different". If the distance between a pair of images is smaller than the threshold, that is a false-alarm (since the images are all different). We use 50 different thresholds that linearly span the full range of distances, in order to examine the full range of outcomes.

From Fig. 10.3 (top), we see that for upright faces, for any given threshold, large, coarse features have higher accuracy (correct-rejection rate) than the small, fine features. This is somewhat counter-intuitive, as one might imagine that fine features would be better for within-category discrimination. The suitability of large, coarse features for both detection and identification is discussed further in Section 11.1.

More importantly, we see that for any given threshold, the FIE for large, coarse features is larger than for small, fine features (Fig. 10.3 bottom). Even when the thresholds for both types of features can be independently chosen so that accuracy for upright faces is matched (Fig. 10.4), the differential FIE still exists.

10.2 Chapter summary

We have shown the that differential FIE is easily explained using our model by comparing large, coarse features ("face-like" processing) and small, fine features ("object-like" processing). Importantly, we give a step-by-step account of the effects of inversion, going from individual unit responses, to distance between faces, to discrimination accuracies.

Other (similar) models have previously also shown the FIE (e.g. Zhang & Cottrell 2004) and differential FIE (e.g. Jiang et al. 2006). The key contribution of our model in this chapter is that we have demonstrated the mechanistic relationship between holism and inversion. "Holistic" (large, coarse) features experience a larger FIE than "non-holistic" (small, fine) features. This is unlike the work of Jiang et al. (2006) in which the differential FIE was accounted for by differential tuning width for face-tuned versus object-tuned units; no link to holism was established.

In previous chapters, we focused on holism (particularly the CFE). We then linked holism to inversion in this chapter. In the next chapter, we proceed to bridge the large gap between holism and configural/face-space/norm-based processing.

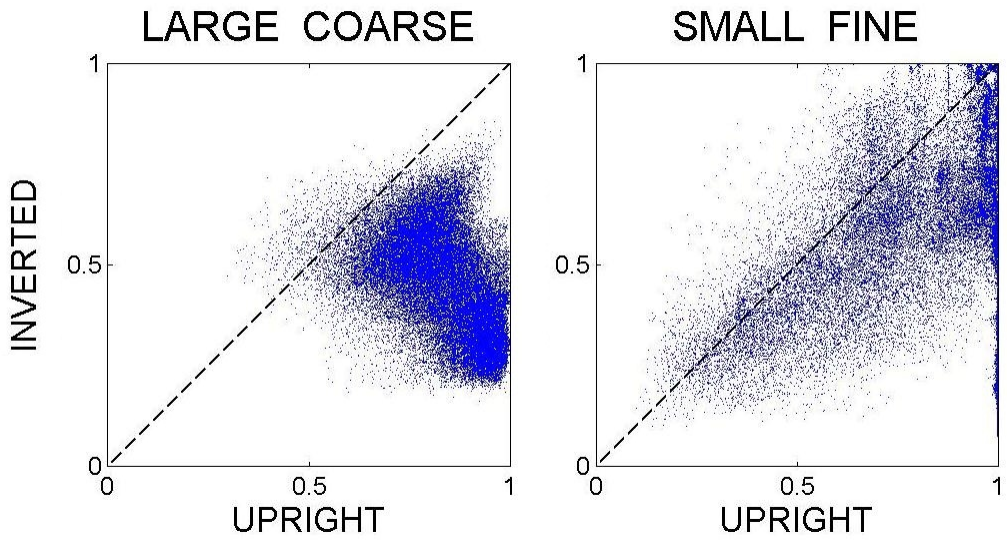


Figure 10.1. Scatter-plot of responses to inverted versus upright faces, for large, coarse features (left) and small, fine features (right).

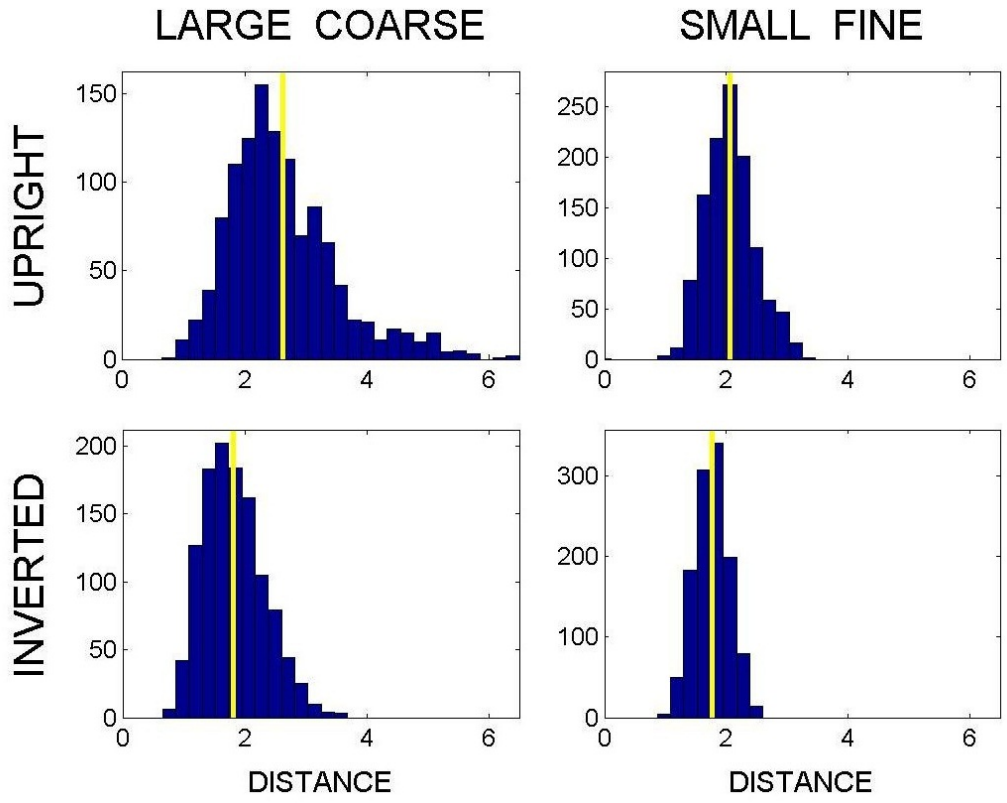


Figure 10.2. Histograms of distance between pairs of faces. Top row: upright faces. Bottom row: inverted faces. Left column: large, coarse features. Right column: small, fine features. Yellow line indicates mean of distribution.

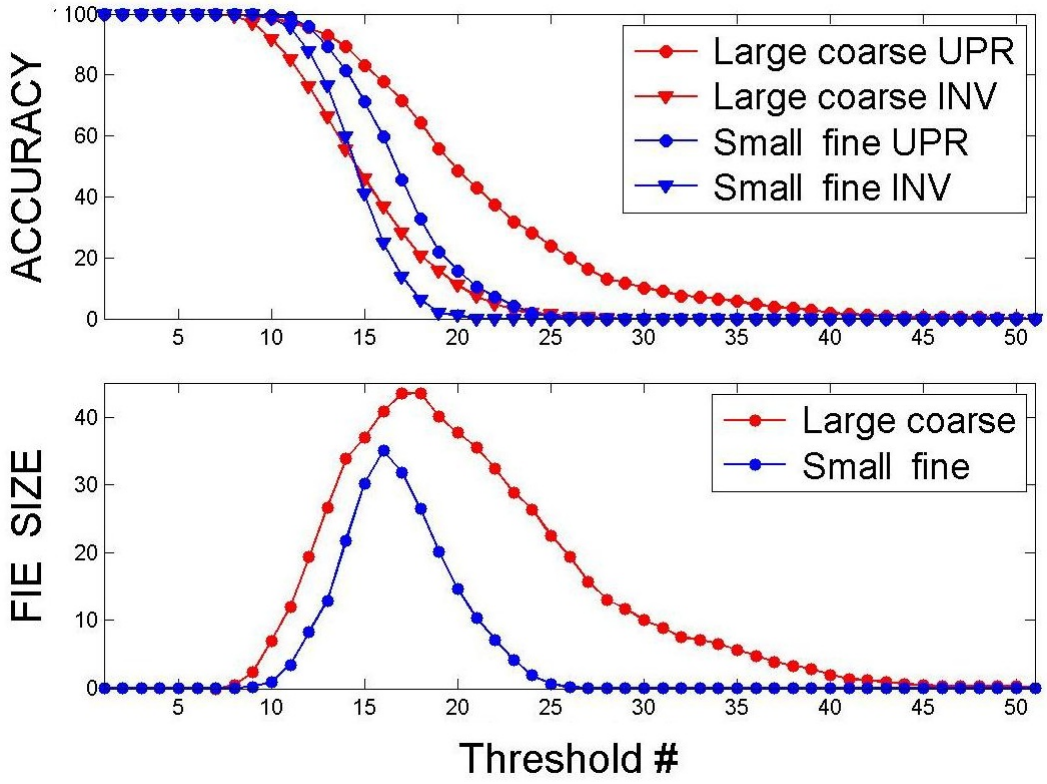


Figure 10.3. Accuracy (top) and FIE size (bottom) for each of 50 linearly spaced thresholds. FIE size is defined as upright accuracy minus inverted accuracy. UPR: upright. INV: inverted.

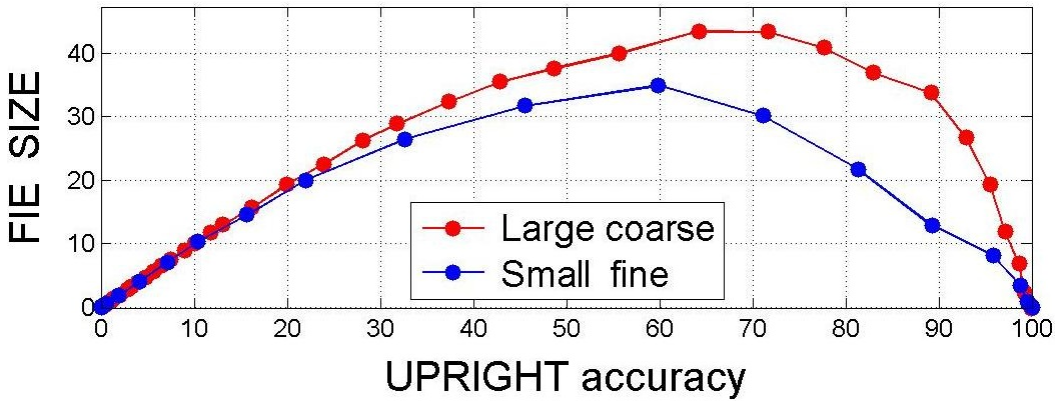


Figure 10.4. FIE size versus accuracy for upright faces.

Chapter 11: Holism and Beyond

Chapter abstract

In this chapter, we attempt to bridge the gap between holistic processing and configural/face-space/norm-based processing. We first show that (surprisingly), large, coarse features are more sensitive than small, fine features to individual identity. We then show this to be the case for second-order configural changes also. Furthermore, our model replicates the ramp-shaped opponent coding for second-order configural changes found in neurons in the macaque middle face patch. Finally, we show that our model replicates some signatures of norm-based coding during adaptation. Crucially, all of these findings were made using our model without any changes, suggesting that all of these aspects of face processing may arise implicitly from large, coarse features rather than through explicit and specialized mechanisms for second-order configuration and norm-based coding.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 11: Holism and beyond

One of the major issues in face processing research is the disconnect between different aspects of face processing. The relationship between holistic processing and inversion is already not well understood. The same is true for the relationship between holistic and configural processing, even though these are sometimes lumped together. The biggest disconnect is between holistic processing and face-space/norm-based coding (McKone 2009).

In this chapter, we attempt to link all of these aspects of face processing using our model. It is important to remember that our model was modified from a model of object processing with the sole aim of replicating the Composite Face Effect (CFE). It is one thing to design a model that just replicates one thing, but a useful model should be able to do more than that. A test of the usefulness of the model as a model of general face processing is therefore whether it can account for phenomena that it was not designed to. We aim to show this here.

11.1 Relating holism to detection and identification

We first examine the issue of detection versus identification. As discussion in Section 3.5, models of face processing disagree over whether holism is linked to detection or to identification. We claim that holism is related to both detection and identification, by showing that even large, coarse features can easily discriminate individual faces.

This claim is also made by a similar model (Riesenhuber & Poggio 2003, Fig. 111.7). The logic is simple. Because units are tuned to faces, non-faces elicit weak responses. At the same time, different faces elicit different responses. Using a population code, individual faces can thus be easily identified. However, the model of Riesenhuber & Poggio (2003) is different from ours; the most salient difference is our usage of large, coarse features. It is unclear if such features are discriminative enough; we will show that they do.

We examined the responses of the 1000 large, coarse templates to 50 different individual faces. (These faces were not the same as the 50 faces from which the templates were extracted.) To gauge the discriminability of each feature (i.e. template), we calculated the difference between maximum and minimum responses to the 50 faces. Fig. 11.1 shows that this difference can be quite substantial. For upright faces, the mean difference is 0.28, which is 32% of the overall mean response (0.87). In other words, large coarse templates can easily discriminate between individual faces, because these faces can elicit quite different responses. Similar to the results presented in Chapter 10 (FIE), the discrimination for inverted faces is worse than for upright faces (Fig. 11.1 bottom).

However, is this discriminability linked to holism, or can any set of templates perform discrimination equally well? From Fig. 11.2, we see that for small, fine templates, the difference between maximum and minimum responses to the 50 faces is small (mean difference of 0.16, compared to 0.28 for large, coarse templates). In fact, we can see that over 400 (out of 1000 total) features respond essentially the same to all the 50 faces (i.e. difference of 0 between max

and min response). This is not because small, fine templates respond poorly to the 50 faces. Their overall mean response is 0.91, which is similar to the 0.87 for the large, coarse templates.

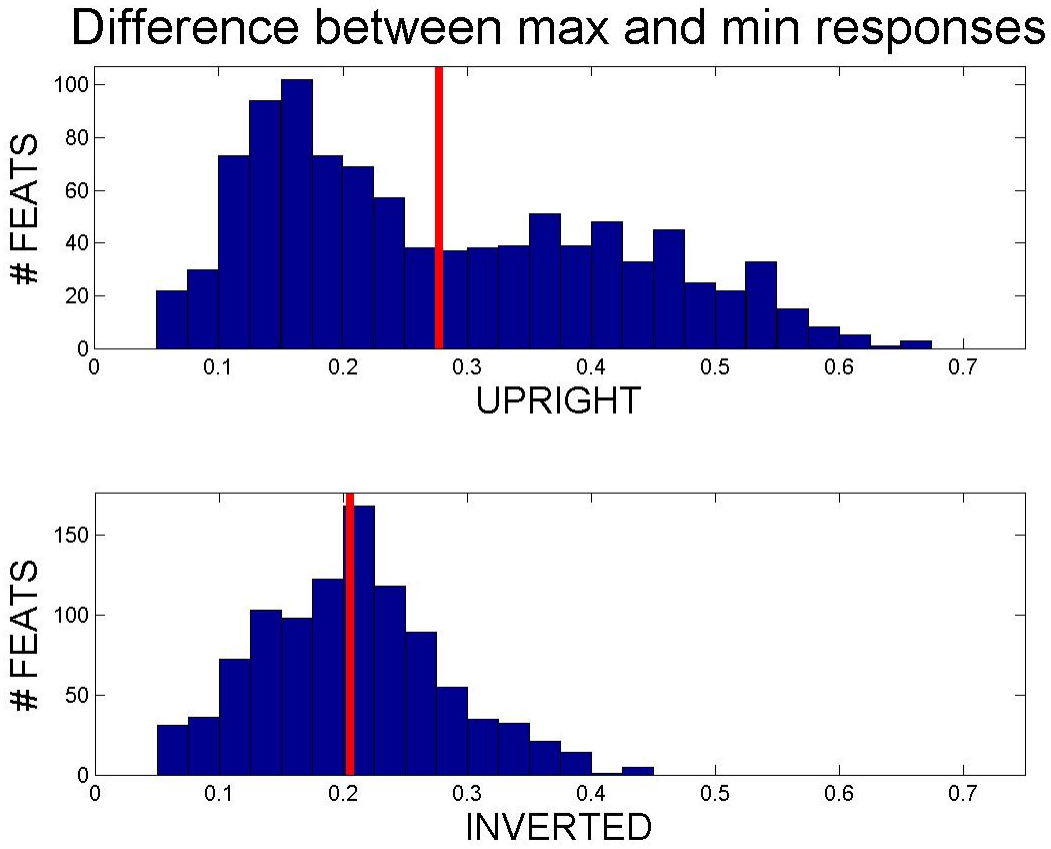


Figure 11.1. Histogram of differences between maximum and minimum responses for large, coarse features. Red line indicates mean. Top: upright faces (mean 0.28). Bottom: inverted faces (mean 0.21).

Why are small, fine templates less discriminative than large, coarse ones? This seems counter-intuitive, and also contrary to at least one other model (e.g. Zhang & Cottrell 2004, 2006). One possibility, specific to our model, is that because the features are position and scale invariant (i.e. the max in the C2 layer is taken over all positions and scales), many "false positives" contribute to the resulting C2 response. For the large, coarse, templates, because of their large size, there are fewer positions to take the max over, so there are fewer "false positives". Further work is needed to determine if this is the true reason.

So, large, coarse templates can support identification. What about detection? Since the set of non-face classes is infinite, rather than compare responses to faces versus some small, arbitrary set of non-faces, we look at faces versus inverted faces. Inverted faces share many physical characteristics with upright faces, and in fact their frequency spectra are identical, except for a phase shift. For large, coarse templates, the mean overall response to inverted faces is 0.47, compared to 0.87 for their upright counterparts. One might then infer than in general, non-faces

would elicit even smaller responses. Thus, face detection can easily be achieved. (Of course, one can always deliberately construct sets of face-like stimuli that may elicit relatively high responses and fool these templates.)

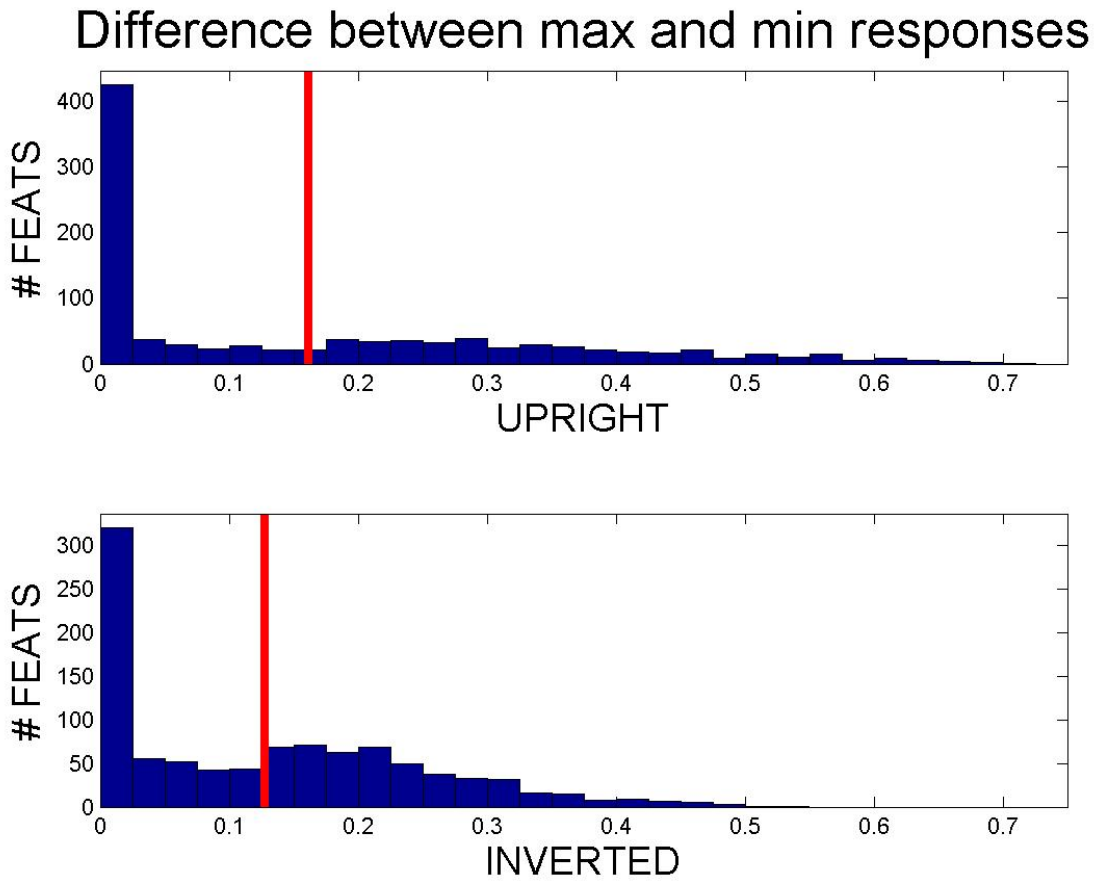


Figure 11.2. Histogram of differences between maximum and minimum responses for small, fine features. Red line indicates mean. Top: upright faces (mean 0.16). Bottom: inverted faces (mean 0.13).

11.2 Implicit coding of second-order configuration

Since large, coarse templates can discriminate between individual faces, could they also be sensitive to configural changes, in particular? The answer is yes. In other words, holistic and configural processing are one and the same thing.

As a proof-of-concept, we created a single large, coarse model unit that is maximally tuned to an "average" cartoon face (Fig. 11.3 middle face). (We return to our regular large, coarse templates in the next section.) Fig. 11.4 (left) shows the response of this unit when the eyes and eyebrows are horizontally or vertically shifted.

From Fig. 11.4 (left), we see that this unit is sensitive to second-order configural changes in eye separation (horizontal change) and eye height (vertical change). Crucially, this sensitivity is implicit, because there were no mechanisms for explicitly measuring eye separation or eye height. This sensitivity is much reduced for inverted faces, as has been found empirically (e.g. Tanaka & Sengco 1997, Freire et al. 2000, Le Grand et al. 2001). Importantly, this reduced sensitivity arises simply from the fact that responses are generally lower, and not because some mechanism responsible for "configural processing" has been disrupted by inversion. The previous section has already shown that small, fine templates are less discriminative, so this control is not shown here.

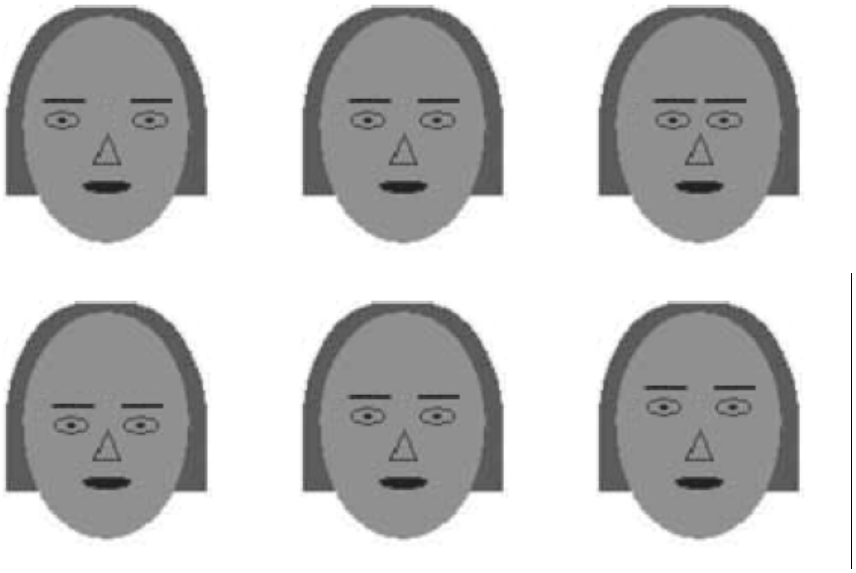


Figure 11.3. Cartoon faces that differ in second-order configuration (i.e. distance between parts). Top row: horizontal distance between eyes. Bottom row: vertical distance between eyes and nose. Faces were adapted from Freiwald et al. (2009).

Interestingly, although the vertical and horizontal changes were of the same amount in pixels, sensitivity to vertical changes was greater. We believe that this is simply due to the fact that faces contain more horizontal contrast energy (already clearly apparent in V1-like responses, see Fig. 11.4 right), rather than more complicated reasons (e.g. Dakin & Watt 2009, Goffaux & Dakin 2010). As a result of greater sensitivity to vertical changes, we also find a larger effect of inversion on vertical than horizontal changes, as reported by Goffaux & Rossion (2007).

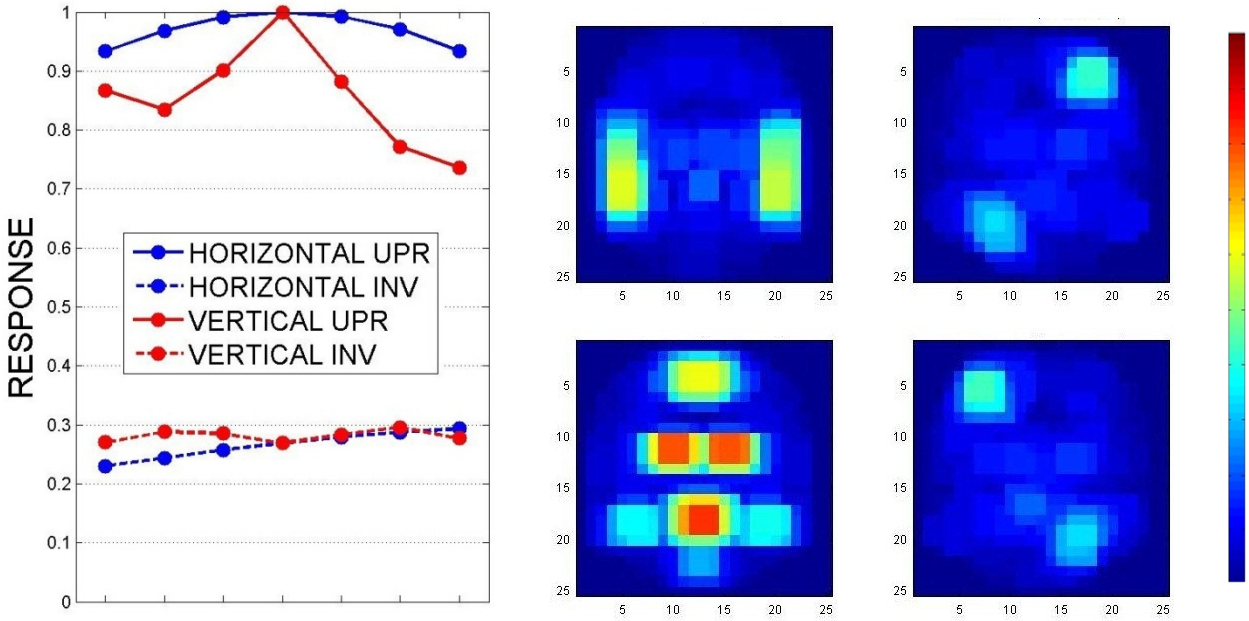


Figure 11.4. Left: responses of one model unit to horizontal (blue) and vertical (red) second-order changes, for upright (solid) and inverted (dashed) faces. Right: C1 (complex-cell-like) response maps to a cartoon face (Fig. 11.3 middle). Orientation channels are (clockwise, from top left) vertical (0°), oblique (-45°), oblique (+45°), horizontal (90°). Low responses in blue, high responses in red.

11.3 Ramp-shaped opponent coding

Freiwald et al. (2009) recorded responses from neurons in the "middle face patch" in the temporal lobe of macaque monkeys, and found ramp-shaped (monotonic, not necessarily linear) tuning for second-order configural changes, consistent with the "opponent coding" theory of norm-based coding (Rhodes & Jeffery 2006). Our model qualitatively replicates their results.

Fig. 11.5 shows the responses of two example large, coarse features to the cartoon faces shown in Fig. 11.3 (top row). Changes in eye-spacing produce monotonic, ramp-shaped tuning curves, like those found by Freiwald et al. (2009).

Over the population of 1000 large, coarse templates, we find tuning properties that are remarkably similar to those found by Freiwald et al. (2009). As evidence for opponent-coding, it was found that most tuning curves had maxima and minima at the extreme feature values (Fig. 11.6 left). This same property was found for our large, coarse templates (Fig. 11.6 middle), whereas this property was less strong for the small fine templates (Fig. 11.6 right).

Ramp-shaped opponent-coding was further evidenced by the finding that the variability in response was larger for the extremal feature values (because the responses can be maxima or minima), compared to the "average" value (Freiwald et al. 2009, Fig. 4c). Again, this property was found for large, coarse templates (Fig. 11.7 top left) more than for small, fine templates (Fig.

11.7 top right). Furthermore, the minimal response was predominantly found at the extreme opposite of the feature value that gave the maximal response (Fig. 11.7 bottom row). This was true 57% of the time for large, coarse templates (67% for middle face patch neurons), compared to 47% for small, fine templates.

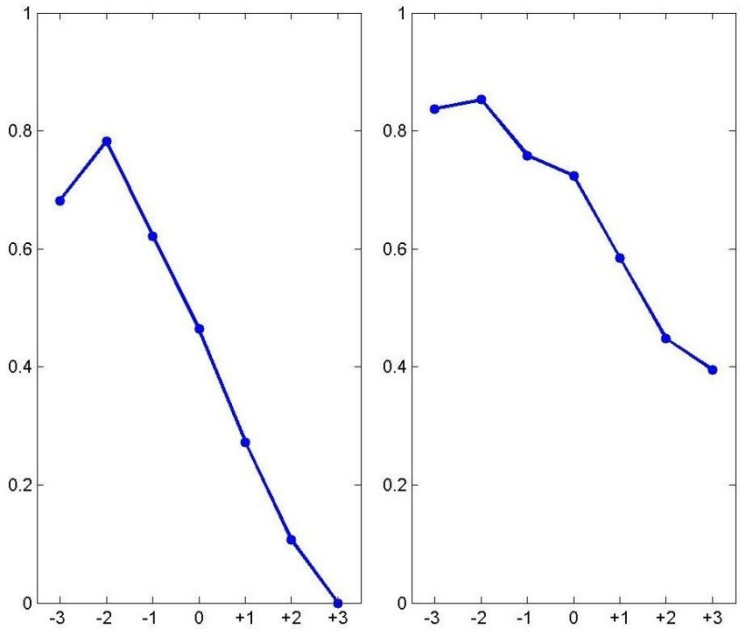


Figure 11.5. Responses of two model units to faces that differ in eye separation. Vertical axis: normalized response. Horizontal axis: eye separation (ordinal units). Faces are depicted in Fig. 11.3, top row (-3: top left. 0: top center. +3: top right).

Altogether, there is good evidence that our large, coarse templates share similar properties with neurons in the middle face patch, in terms of opponent coding for second-order configuration. Importantly, small, fine templates were less similar than large, coarse templates to these neurons. Thus, we have established a link between holistic processing and opponent coding. Crucially, however, our model does not have any explicit mechanisms designed to produce such opponent coding. So, how does this come about?

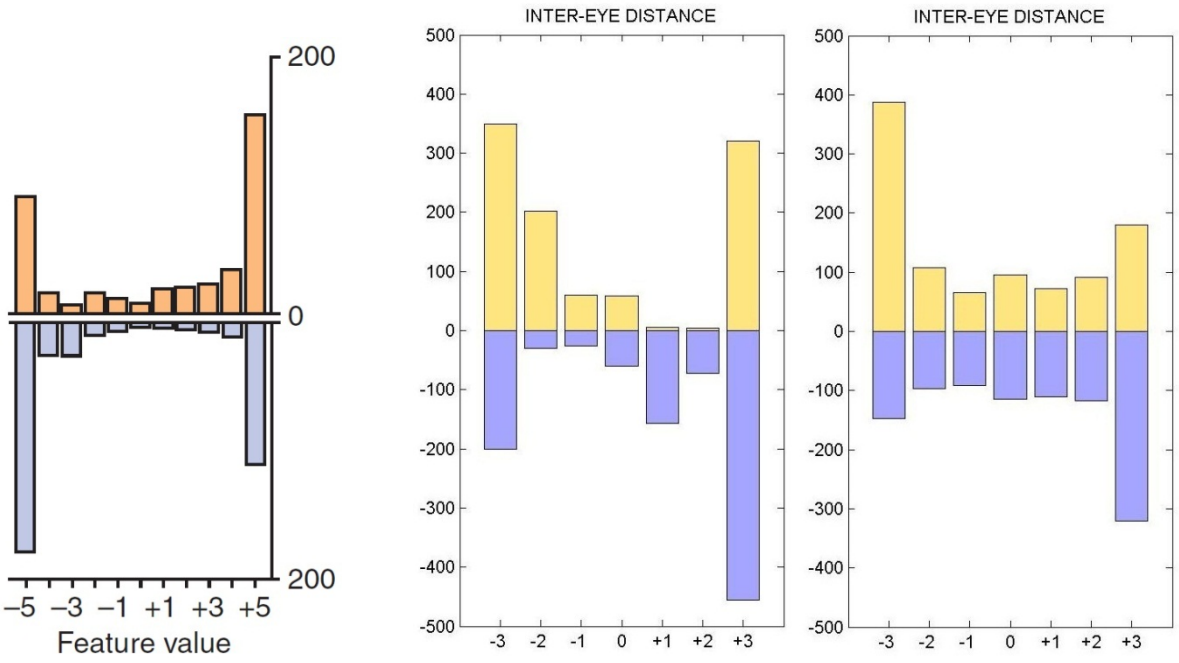


Figure 11.6. Number of neurons or model units with maxima (orange/beige) or minima (blue) at each feature value. Left: neurons (from Freiwald et al. 2009). Middle: large, coarse templates. Right: small, fine templates. **Note:** we used stimuli similar to Freiwald et al. (2009), but did not have their exact stimulus set.

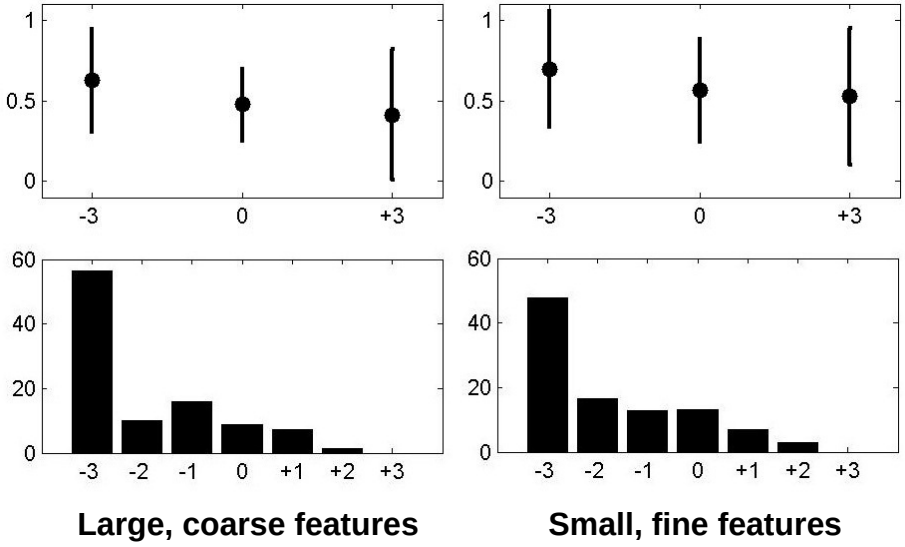


Figure 11.7. Top row: standard deviation in response over all features for extremal and "average" feature values (Left: large, coarse. Right: small, fine). Bottom row: percentage of features for which the minimal response was found at a given feature value (Left: large, coarse templates. Right: small, fine templates). **Note (bottom row only):** following Freiwald et al. (2009) Fig. 4d, feature values were flipped when necessary, so that +3 corresponded to maximum response.

There are two aspects to opponent coding: ramp-shaped tuning and opponency. We examine how our model produces either of these in turn. The ramp-shaped tuning arises simply because the physicals changes are relatively small, compared to the space of all possible changes that could be made. As illustrated in Fig. 11.8, for small changes in the input space, gaussian tuning (used by our model) is monotonic and relatively linear for most local sections of the tuning curve (Fig. 11.8, black circle). Only occasionally will the changes span the "hump" sections of the turning curve and produce bell-shaped tuning (Fig. 11.8, black rectangle). In more technical terms, it is only when the direction of change is orthogonal to the direction of the gradient, will the local section of the overall tuning curve be bell-shaped. It should be noted, however, that this is only true for relative small changes (i.e. small sections of the tuning curve). We predict that for the neurons found by Freiwald et al. (2009) to have ramp-shaped tuning, spanning a larger region of the space will produce bell-shaped tuning rather than ramp-shaped tuning. This could be achieved by showing morphs from an object to a face to the "anti-object" that is "on the opposite side", for instance.

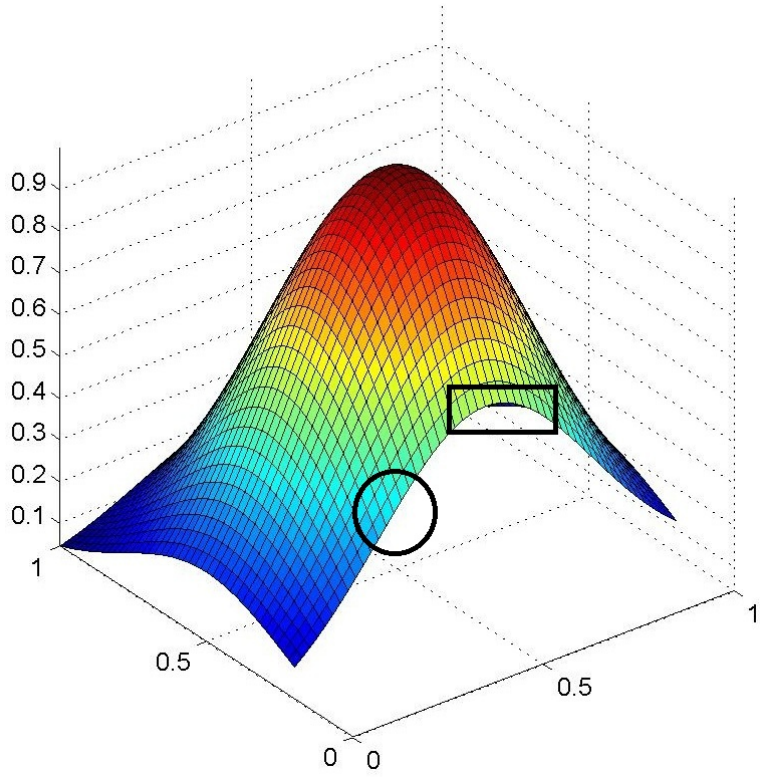


Fig 11.8. Gaussian tuning curve, cutaway for easier visualization. Maximum output (1.0 on the vertical axis) occurs for input values of [0.5, 0.5]. For small changes in the input, changes in the output are mostly ramp-shaped (black circle). Changes are only bell-shaped under certain circumstances (black rectangle).

Apart from ramp-shaped tuning, the second aspect of opponent coding is opponency: the notion that for each metric feature such as eye separation, there exists two populations of neurons, each with slopes of opposite signs (Fig. 11.9 left). The norm (average face) is implicit rather than explicit, but is given a special status, encoded by the equal firing of both neuronal populations.

How does this come about in our model, especially since the model per se accords no special status to the average face? The answer is illustrated in Fig. 11.9 (right).

For any set of test faces that differ along some metric feature (squares) and are centered on the "average face" (filled square), arbitrary faces (blue circles) will generally be "positive" (upper left) or "negative" (lower right) with respect to the average face in roughly equal proportion. This stems from the physical facial properties, which are normally (or at least symmetrically) distributed (Shepherd et al. 1977, Bruce et al. 1994, 1995). As such, the distances from the arbitrary faces (blue circles) to the test faces (squares) will be roughly balanced about the average face, similar to the depiction in Fig. 11.9 (left). Importantly, there is no explicit opponency mechanism, and the balancing arises naturally from the roughly normal distribution of the physical facial properties.

This mental model addresses one of the common reservations about norm-based coding: different studies all claim to support the special status of the face norm, but this norm can be quite different in each study (Tsao & Freiwald 2006)! For example, Leopold et al. (2001) use only male faces, so their norm is unambiguously male. Another study might investigate neural coding of gender, so its norm would be a gender-ambiguous face. How can the norm be accorded a special status if there is no agreement on what the norm is?

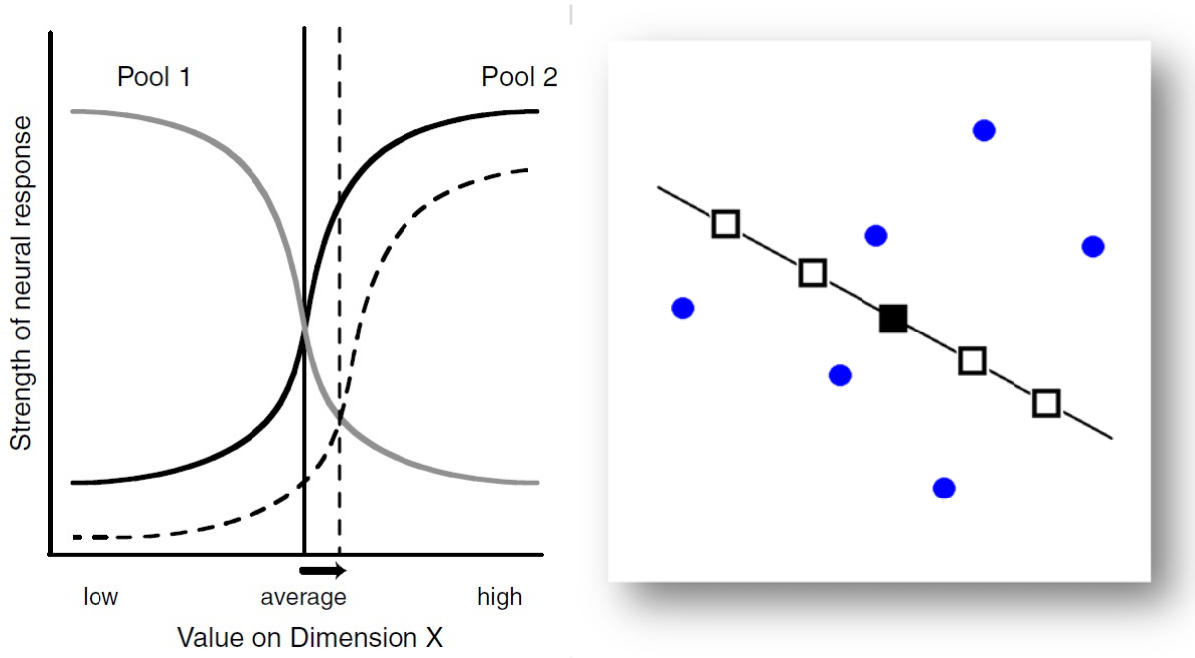


Figure 11.9. Left: illustration of opponent coding. Right: scheme for opponent coding without explicit mechanisms. Squares: faces that vary along some arbitrary dimension. Filled square: "average" face. Blue circles: arbitrary faces (exemplars from which distances are calculated, simulating responses of neurons tuned to arbitrary faces). Left: figure from Rhodes & Jeffery (2006). See p.16 for copyright notice.

Our model, which derives its large, coarse templates by randomly sampling patches from some arbitrary set of faces, contends that it is not the norm per se that has some special status. Rather, because of the physical properties of faces, the approximate average along any arbitrary dimension will induce a roughly symmetric distribution of distances on either side of this approximate average. Importantly, however, this model predicts that the norm does not need to be average along all other dimensions also. (This would be needed if the "truly average face" had a special status). For example, if the dimension of interest is eye separation, but this eye separation is varied for a very old, masculine face with atypical eye height and face width, then evidence for norm-coding can still be found, e.g. from adaptation studies. Of course, evidence may be weaker, e.g. weak adaptation, because the number of neurons that would be strongly activated by these faces may be small. Consequently, the model predicts that for faces that vary in multiple dimensions (e.g. variation along 2 dimensions, each with N steps), there will be an interaction between the dimensions. This is not a trivial prediction. One could imagine models where eye separation is explicitly calculated, and thus neurons that code for this are completely independent from dimensions such as eye height or mouth width.

The norm-based opponent model is sometimes contrasted to the exemplar-based "multi-channel" model (Fig. 11.10 right). We view this as a false dichotomy that is based on quantitative rather than qualitative differences. To see this, if the neurons in the multi-channel model (Fig. 11.10 right) are as broadly tuned as those in the opponent model – and the electrophysiological evidence suggests that face cells are indeed broadly tuned – then there is little difference between the two models. Neurons that are tuned to non-extremal faces will respond roughly equally strongly to all faces, making them effectively neutral to adaptation effects. Thus, a broadly-tuned multi-channel model is effectively indistinguishable from the opponent model. In fact, it is more parsimonious, because it makes no assumptions about the special status of the average feature value or average face, and can account for the fact that different studies show evidence for different norms (as explained earlier).

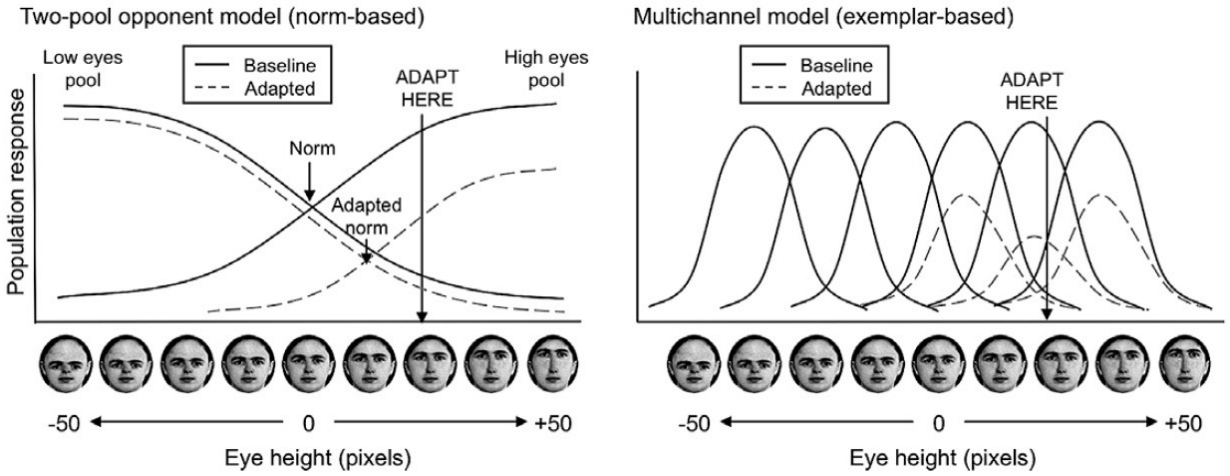


Figure 11.10. Face space models. Left: opponent-coding (norm-based) model. Right: multi-channel (exemplar-based) model. Figure from Susilo et al. (2010). See p.16 for copyright notice.

11.4 Adaptation and norm-based coding

The adaptation paradigm is often used to investigate norm-based coding. The idea, illustrated in Fig. 11.9 (left), is that after prolonged exposure to a face with some non-average feature value, neurons that code for this feature experience "fatigue" and fire less strongly. As such, the point of balance between opponent pools (which is assumed to code for the norm) becomes shifted towards the adapted face. Perceptually, subjects report that after adaptation, the adapted face seems more average, and the actual average face seems more extreme.

Our model also shows this behavior. We implemented adaptation straightforwardly by attenuation each feature's response proportional to its response to some adapter face. We then tested these post-adaptation features to a set of faces that varied metrically in eye separation. There were 7 such faces, labeled -3 to +3. To measure the "perception" experienced by this set of features, we performed linear regression for each feature to find the slope and offset that would allow each feature's response to predict the face label (e.g. +2, -1, etc.). Many features had slopes close to 0, and were thus non-informative. Therefore, out of 1000 features, we only analyzed the 200 features with the most negative and positive slopes. Importantly, at no point do we introduce any explicit coding of the "norm" (face 0), nor accord it special status.

First, we verify that before adaptation, features with either positive or negative slopes indeed "perceive" the faces veridically, according to their actual labels (black lines, Fig. 11.11 left and right). We then examined the "perceived" face identities after adaptation to either the -3 face (blue lines) or the +3 face (red lines). Consistent with the opponent coding model, features with positive slopes were more affected by adaptation to +3 than -3 (Fig. 11.11 left). Conversely, features with negative slopes were more affected by adaptation to -3 than +3 (Fig. 11.11 right).

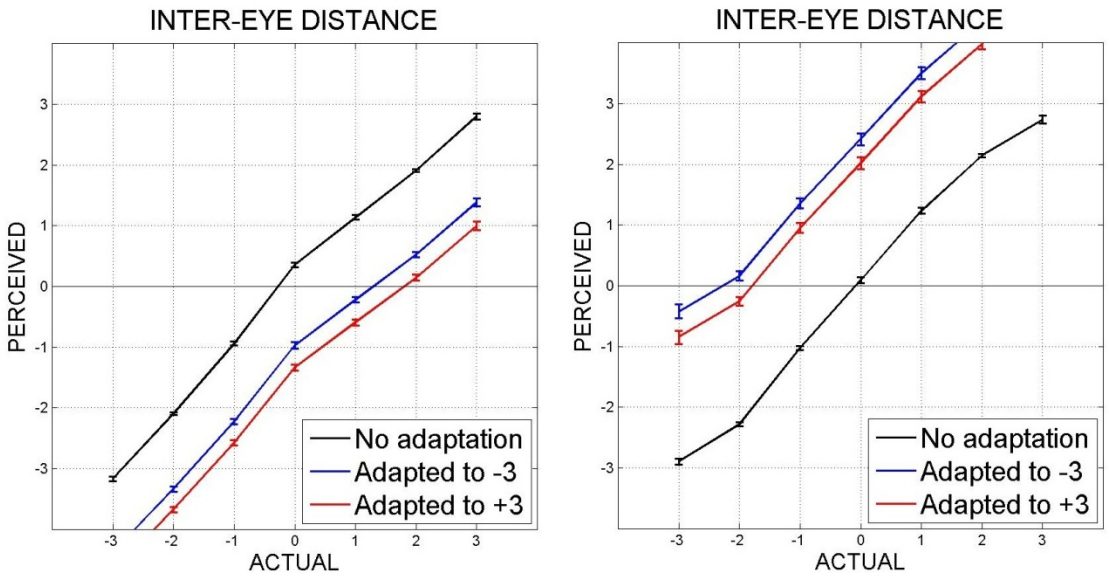


Figure 11.11. Effects of adaptation on face "perception" on opponent unit pools. Left: model units with positive slopes. Right: model units with negative slopes. Note: "slope" indicates whether responses increase or decrease with face number (Fig. 11.9 left), not the slope of the lines shown in this figure.

Looking at both the positive- and negative-slope features together, we see that adaptation to -3 causes a relative shift towards -3, compared to adaptation to +3 (Fig. 11.12 left). Note that in absolute terms, adaptation to +3 also shifted the (red) curve left, relative to veridical perception (the y=x diagonal). This is anomalous. There are two possible reasons for this. Firstly, we picked the '0' face arbitrarily (copying the '0' face in Freiwald et al. 2009; note the asymmetries in Fig. 4a and 4b, suggesting that this face is not quite "average"), so that there is no guarantee that it is actually "average" in any sense. Secondly, our features were extracted from 50 arbitrary faces, which is a relatively small number, and may not reflect the balanced, symmetric distribution of properties found in more realistic, broader sets of faces.

In Figs. 11.6 and 11.7, we showed links between large, coarse templates and opponent coding. Is this also true for adaptation? Fig. 11.12 (right) shows that for small, fine templates, the effects of adaptation are not significantly different for adaptation to -3 and +3. This is also true even when features with positive and negative slopes are analyzed separately (Fig. 11.13).

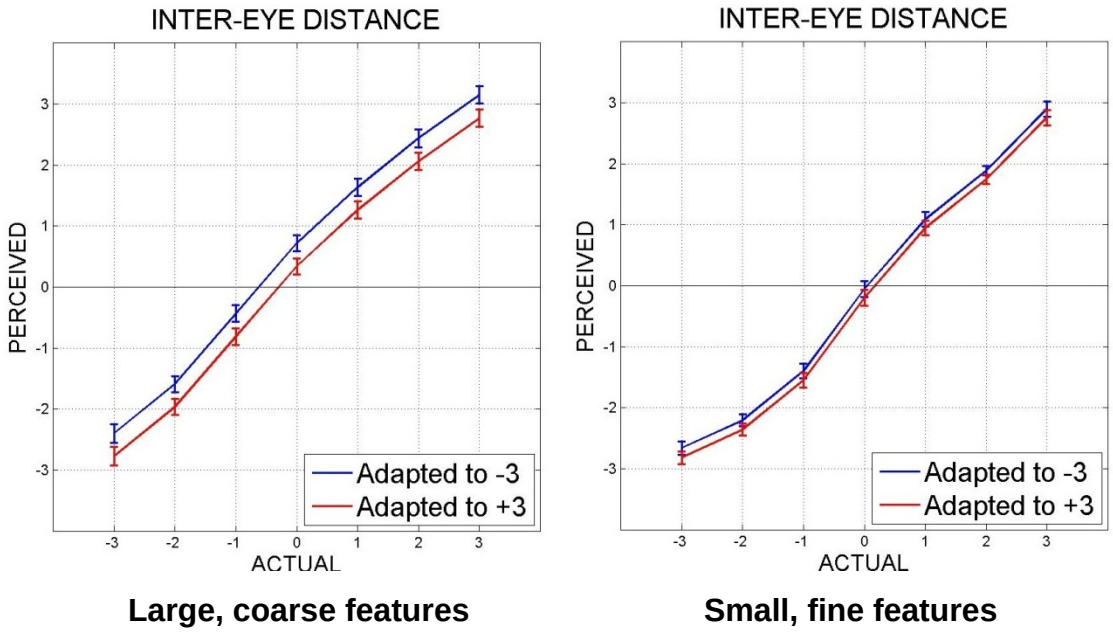


Figure 11.12. Effects of adaptation. Combined results for neurons with positive and negative slopes. Left: large, coarse features. Right: small, fine features.

The quantitative difference between adaptation effects for large, coarse templates and small, fine templates is not due to insufficient adaptation. When the parameter controlling adaptation strength is 10 times as strong as before (leading to behaviorally implausible levels of adaptation, shown in Fig. 11.14), the difference between adaptation to -3 and +3 is still apparent for large, coarse features (Fig. 11.14 left), but not for small, fine features (Fig. 11.14 right).

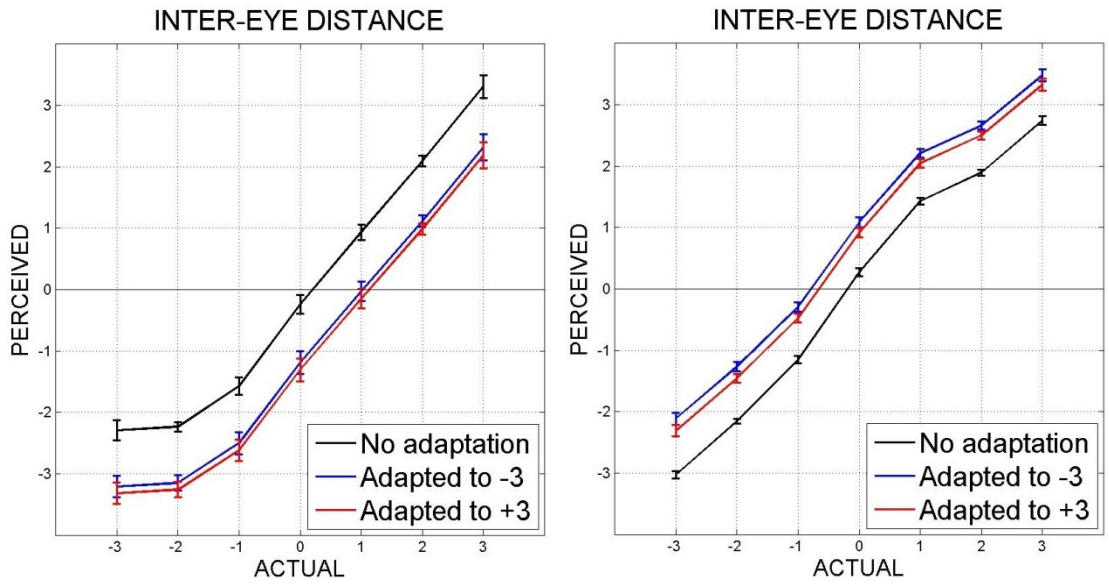


Figure 11.13. Effects of adaptation for small, fine templates. Left: model units with positive slopes. Right: model units with negative slopes.

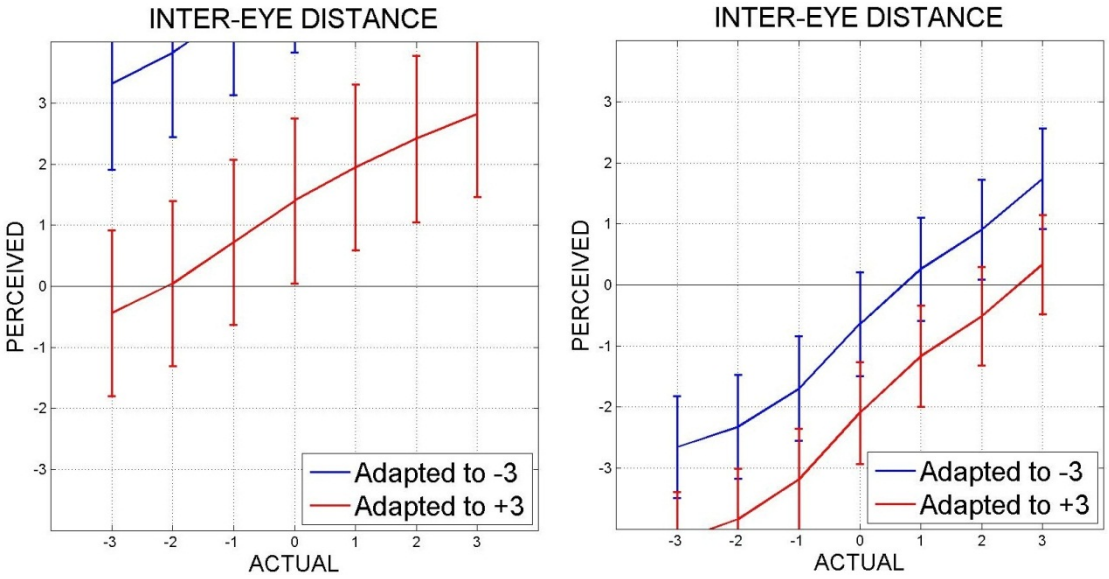


Figure 11.14. Effects on adaptation, with adaptation strength 10 times of that used to produce Fig. 11.12. Left: large, coarse features. Right: small, fine features.

Finally, it has been found that adaptation for upright and inverted faces can be independent (Rhodes et al. 2004, Susilo et al. 2010). Our model easily explains this. Similar to what was discussed in Chapter 10 for the FIE, since the correlation of responses to upright and inverted faces is negative, the features that respond most strongly to upright faces are those that respond

the weakest to inverted faces (Fig. 11.15 left). Therefore, adaptation to an upright face affects these features the most, while adaptation to an inverted face affects them least. In other words, independent adaptation to upright and inverted faces is not because there are separate neuronal populations coding for these. (In our model, all templates are derived from upright faces, and virtually all respond more strongly to upright than inverted faces). On a related note, the fact that the correlation is, instead, positive for small, fine features (Fig. 11.15 right) makes the prediction that adaptation to upright and inverted images is not independent for "object-like" processing.

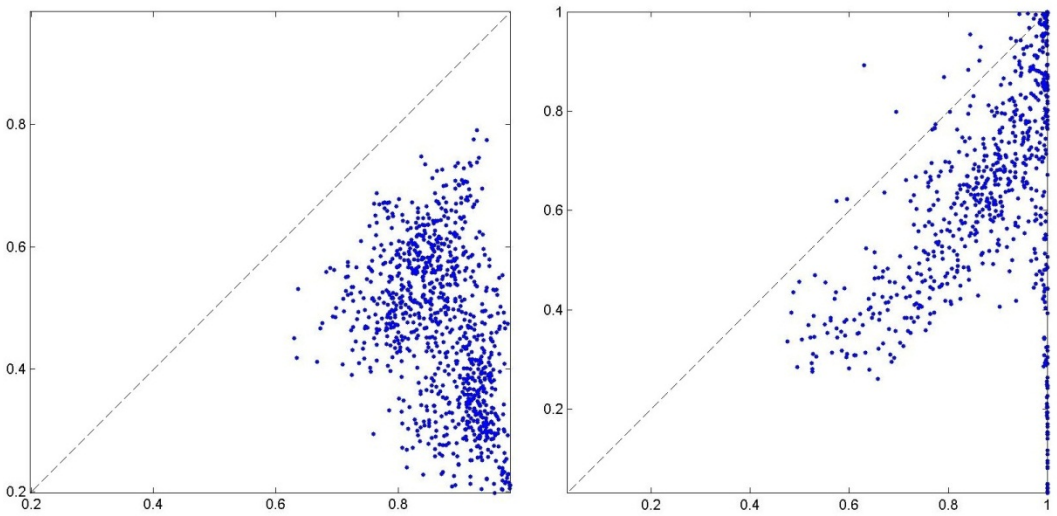


Figure 11.15. Responses to inverted (y-axis) vs. upright faces (x-axis). Each point represents the mean response (averaged over 50 faces) of one model unit. Points below the diagonal show inversion effects. Left: large, coarse features (r = -0.37). Right: small, fine features (r = +0.57).

We end off by recapping the conditions required to produce the effects attributed to opponent, norm-based coding, even in a model that does not explicitly have such coding. First, the physical properties of the stimuli must be symmetrically distributed about the mean. Bodies, for example, may be suitable, but letters are probably not. Second, the features must be sensitive to the relevant stimulus changes, such as changes in body height/width ratio. Since many of these changes are "configural" in the sense that they involve relationships between different parts of the stimulus, only "holistic" features may show opponent coding for such changes (dimensions). A related prediction is that for faces, small fine templates may exhibit opponent coding for "local" changes such as nose width, but not for "configural" changes with as eye separation.

11.5 Chapter summary

We attempted to bridge the gap between holistic processing and configural/face-space/norm-based processing. We first showed that large, coarse features are more sensitive than small, fine

features to individual identity. We then showed this to be the case for second-order configural changes also. Our model also replicated the ramp-shaped opponent coding for second-order configural changes. Finally, we showed that our model replicates some signatures of norm-based coding during adaptation. Crucially, all of these findings were made using our model without any changes, suggesting that all of these aspects of face processing may arise implicitly from large, coarse features rather than through explicit and specialized mechanisms for second-order configuration and norm-based coding.

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 12: Alternative Accounts

Chapter abstract

In this chapter, we compare our account of the CFE with the more widespread, intuitive account. We first show that our account is better in terms of generalizability. Nonetheless, we proceed to compare the predictions from the two accounts. We find that there is some empirical support for our model, but that more targeted experiments need to be conducted. We then proceed to show that the two accounts may actually differ in the "decision-making" stage, rather than the core "holistic processing" stage. Furthermore, we show that our model can actually show characteristics of both accounts under different circumstances. Overall, this chapter illustrates some pitfalls of relying solely on intuitive "mental models" to understand phenomena and make predictions.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 12: Alternative Accounts

So far, the explanations of how our model produces various effects (e.g. misalignment effect, congruency x alignment interaction, and face inversion effect) have mostly hinged on reduced responses of model units to misaligned and inverted faces. Looking at this from the neural and mechanistic perspectives, it seems reasonable that face-tuned units respond less to these "non-standard" face stimuli. Accordingly, this "reduction" account of holistic face processing may also seem reasonable.

However, these effects can also be explained from a more cognitive or psychological perspective. We utilized this more intuitive "influence" account (see Section 12.2) in Chapter 2, where we reviewed the CFE independently of our model. Under this account, misalignment or inversion "disrupt holism" in some unspecified way.

Are these two accounts compatible with each other? Are they really just the same thing described in different ways? If they are different, do they actually make differing predictions that can be tested empirically? This chapter explores these questions.

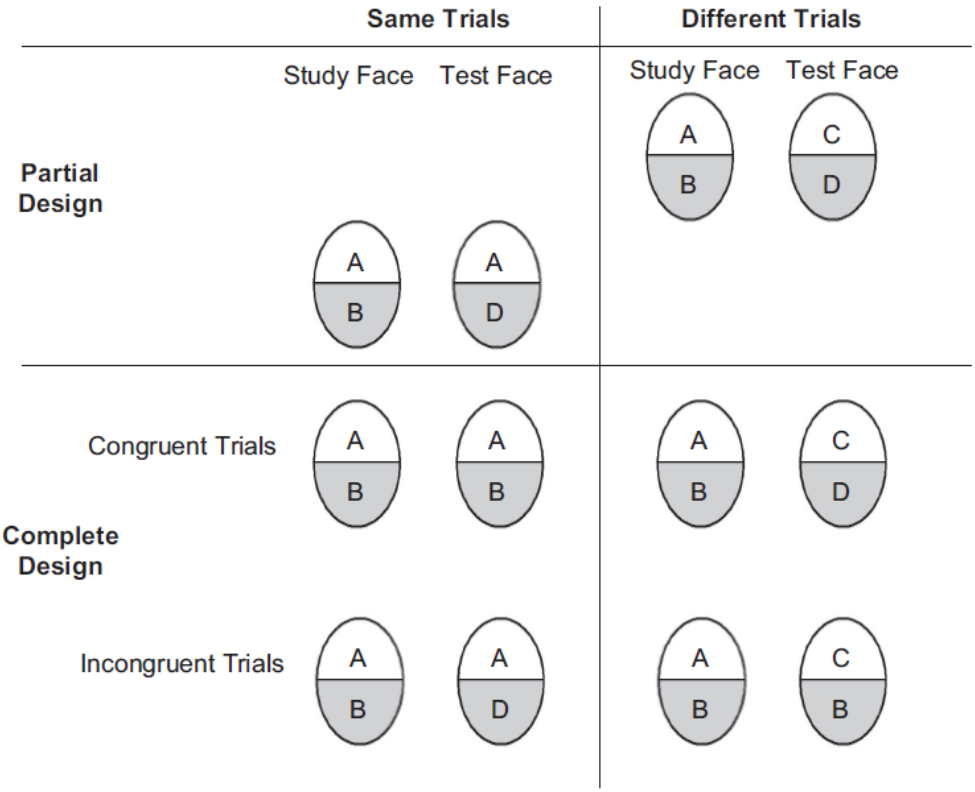


Figure 12.1. Trial types for the CFE "partial" and "complete" designs. Congruent trials are those in which the top and bottom halves are either both same or both different. Note that the "partial" design is a subset of the "complete" design. (Figure reproduced from Cheung et al. 2008. See p.16 for copyright notice.)

12.1 The "reduction" account

We briefly recap the "reduction" account for the misalignment effect here (see Chapter 5 for a more detailed explanation). First, we note that misaligned composites elicit somewhat smaller responses than their aligned counterparts. Secondly, we note that generally speaking, the distance between two vectors (each representing the responses to a composite) becomes smaller when the elements in the vectors become smaller (e.g. due to misalignment). Finally, assuming a relatively constant threshold, smaller distances imply higher hit rates. In short, misaligned composites elicit smaller responses, which lead to smaller distances, which lead to higher hit rates. Thus, the misalignment effect is produced.

But what about holistic processing per se, independent of the effects of misalignment? Because the templates are large, many include portions of both the top and bottom halves. Therefore, unless the bottom halves are completely attenuated due to attentional effects, the bottom halves will contribute to the responses of many units. In the case of incongruent-same trials (see Fig. 12.1), this means that the identical top halves will inevitably be "perceived" to be non-identical by the model (see Section 13.4.4 for more detailed explanation).

12.2 The "influence" account

The "influence" account is not very different from the "reduction" account when it comes to explaining holism per se. Under this account, the bottom halves "influence" perception of the top halves (by unspecified mechanisms). Therefore, in the incongruent-same trials, because the bottom halves are different, the identical top halves are perceived to be non-identical.

However, when it comes to explaining the misalignment effect, the "influence" and "reduction" accounts differ. Under the "influence" account, misalignment "disrupts holism" (again, by unspecified mechanisms). Therefore, the top halves are not influenced by the bottom halves, and are perceived to be identical. Note that this account makes no prediction about the responses to each composite, only the distance (or similarity) between the two composites. The "influence" account, although it has not been explicitly termed as such, is currently the commonly accepted account.

12.3 Comparing accounts: generalizability

A model is only useful if it not only reproduces the phenomenon it is designed to show, but if it can also explain how the phenomenon comes about. By this criterion, the "influence" account is less useful. Unlike the "reduction" account, the "influence" account currently has no explanation (quantitative or qualitative) of plausible neural or computational mechanisms that give rise to holism.

Additionally, a good model should also generalize, i.e. predict (or at least "post-dict") phenomena other than the phenomenon that the model was designed to reproduce. As we have described in the previous chapters, the "reduction" account seems to do a credible job of this.

However, the "influence" account is rather vaguely specified, so it is hard to make predictions. One could say that it predicts that holism is disrupted when composites are manipulated in any way that makes them differ from "normal" faces. This would be consistent with the results for inverted and misaligned composites. However, disregarding the fact that the composites are already quite different from normal faces (gap between the halves; oval crops that exclude hair, ears, and normal face shape), the fact that contrast-reversed faces (clearly not normal) are perceived holistically (see Chapter 8) nullifies this argument.

One might then modify the argument to say that only changes that disrupt the normal first-order configuration (again, this is somewhat poorly specified) will disrupt holism. However, Taubert & Alais (2009) found that vertically shifting the bottom halves (which does not change first-order configuration) produces a misalignment effect, i.e. disrupts holism.

One might then again modify the argument to say that faces must have biologically plausible second-order parameters (which is how Taubert & Alais 2009 interpret their results), otherwise holism is disrupted. This may well be true, but then now the "influence" account becomes less distinguishable from the "reduction" account. According to the "reduction" account, composites that produce reduced responses will consequently produce a higher hit rate. This is not so different from saying that non-biologically plausible changes to second-order configuration (which presumably elicit reduced neural responses) disrupt holism.

Thus, in the end, the two accounts are not so different, but the "influence" account had to be modified and extended post-hoc to remain consistent with the empirical evidence. Crucially, the "influence" account still does not give a mechanistic explanation, nor does it link holism to other important aspects of face processing.

12.4 Comparing accounts: "partial" and "complete" designs

We now return to the CFE for regular, upright composites and examine in more detail if the two accounts really are that similar. We first examine the "partial" design. Since both accounts were "designed" to account for the misalignment effect in "same" trials, we compare the predictions of either account for "different" trials.

As discussed in Section 2.1.1, the intuitive "influence" account makes no predictions for the effect of misalignment on the "different" trials, since it is unclear whether the bottom halves (B and D) will influence the top halves (A and C) to seem "more different" or "less different" (it helps to refer to Fig. 12.1). On the other hand, the "reduction" account makes a clear prediction. Like for the "same" trials, misalignment reduces the responses and distances for the "different" trials. Hence, misalignment should lead to a drop in accuracy (i.e. increased false alarm rate), at least in theory. In practice, it is possible that the reduction in distance is insufficient to cause the distances to fall below the threshold. Section 12.5 examines what actually happens empirically.

We now turn to the "complete" design, and examine the other two conditions not covered in the "partial" design. For congruent-same trials, i.e. two identical composites, both accounts make the

same prediction – as should any reasonable account. In both cases, misalignment should cause little change, since the two composites have identical top and bottom halves.

The incongruent-different trials are the most interesting condition. According to the "influence" account, since the bottom halves are identical, they influence the top halves to seem more similar. Misalignment should therefore lead to a decrease in false alarm rate (or no change, if the effects are subtle). In contrast, the "reduction" account predicts that since the distances are decreased by misalignment, there should be an increase in false alarm rate (or no change, if the effects are subtle).

Now that we have made intuitive predictions for both accounts, we examine the empirical data and model results to see if the predictions are correct. In particular, we look at the congruent-different and incongruent-different conditions, since both accounts make similar predictions for the "same" trials.

12.5 Existing empirical data

For the congruent-different condition, the "influence" account makes no clear prediction. As such, no empirical result would be inconsistent with it. However, if the empirical results clearly favor one outcome over another, then this account cannot explain this. The "reduction" account predicts that the false alarm (FA) rate should increase (or stay the same, due to pragmatic factors such as experimental noise or lack of statistical power). If the FA rate decreases, this account has no good explanation.

What do the empirical results show? We examined all the CFE studies (to the best of our knowledge) that reported the FA rates. Since the widespread intuition is that there is no prediction for the "different" trials, few papers reported FA rates. These results are summarized in Table 12.1. (**Note:** all studies reported CR rates, rather than FA rates. Hence, Table 12.1 lists CR rates. The prediction from the "reduction" account is that CR rates should decrease)

As the results in Table 12.1 indicate, all but one experiment show a decrease in the CR rate, consistent with the "reduction" account's prediction. For the one experiment that does show an increase, this is clearly non-significant (de Heering et al. 2007, Table 2). For the experiments that showed an decrease in CR rate, 2 had non-significant trends, 2 had trends of unreported significance, and 1 had a significant effect. Together, these results are not in any way ironclad proof supporting the "reduction" account, nor do they disprove the "influence" account. However, they do constitute a "proof-of-concept" for the "reduction" account that justifies further investigation involving targeted experiments (see Section 12.7).

Study	Data source	CR aligned	CR misaligned	Trend consistent w/ "reduction"	Signif.
Le Grand et al. (2004)	Fig. 2 & p.764	See source figure		**Y**	p>0.1
de Heering et al. (2007)	Table 1 & p.63	87 (12)	81 (14)	**Y**	p=0.017 [1]
	Table 2 & p.66	90 (9)	92 (11)	**N**	n.s. [2] $F(1,56) < 1$
Robbins & McKone (2007)	Table 7	75 ± 4	70 ± 4	**Y**	--
Cheung et al. (2008)	Appendix C	96	89	**Y**	--
Rossion & Boremanse (2008)	Table 1a & p.6	94 ± 2	89 ± 4	**Y**	p>0.22 [3]

Table 12.1 Correct-rejection (CR) rates (i.e. accuracy for "different" trials) across various studies. Numbers in parentheses indicate standard deviations. Numbers after ± indicate standard errors.

-- indicates significance not reported.
[1] p-value is for main effect of alignment across all age groups tested. No main effect of age, nor interaction between age and alignment was found.
[2] Reported F-value is for main effect of alignment across all age groups tested. No interaction between age and alignment was found.
[3] p-value was for main effect of alignment across all orientation conditions. A t-test for the upright condition was not performed.

We next turn to the incongruent-different condition, used only in the "complete" design. The "influence" account predicts a decrease in FA rate for misalignment, while the "reduction" account predicts an increase. Non-significant changes are consistent with both accounts. Only one study (Cheung et al. 2008) published the raw FA rates (equivalently, correct rejection or CR rate). Here, there was no significant difference in CR rate (93.09% aligned versus 93.10% misaligned), so further experiments need to be conducted (see Section 12.7).

12.6 Closer examination of model predictions

Before we discuss the experiments proposed to specifically validate our model's predictions, let us examine these predictions more closely. We made those predictions by reasoning intuitively from the "reduction" account. Does the model actually even make those predictions?

12.6.1 What does the model actually predict?

In Section 5.3.2, we explicitly stated that smaller responses generally lead to smaller distances. This is not always true. Furthermore, while both Euclidean distance and Pearson correlation are able to replicate the misalignment effect (Section 5.4.4), the congruency effect (results not shown) and the (congruency x alignment) interaction (results not shown), we have not yet examined if both distance metrics have the same effect on "different" trials per se.

As Fig. 12.2 (top) shows, for regular (full spectrum) faces, using Euclidean distance, the false alarm rate increases as predicted. This is true for both congruent and incongruent trials. Using Pearson correlation as the distance metric, qualitatively similar effects are found (results not shown). In other words, the predictions for the "reduction" account from intuitive reasoning and from model results match.

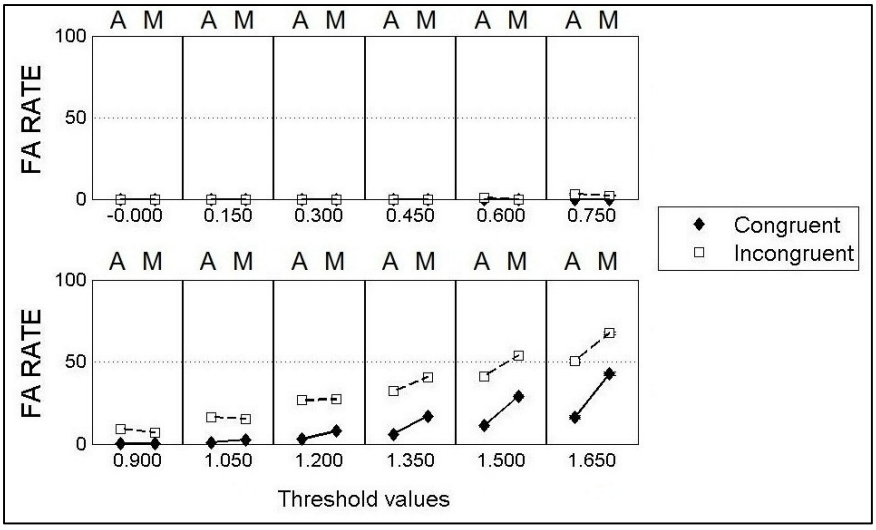
(The relevance of the other spatial frequency conditions, i.e. LSF and HSF, will become clear in Section 12.6.2).

12.6.2 Does our model really implement the "reduction" account?

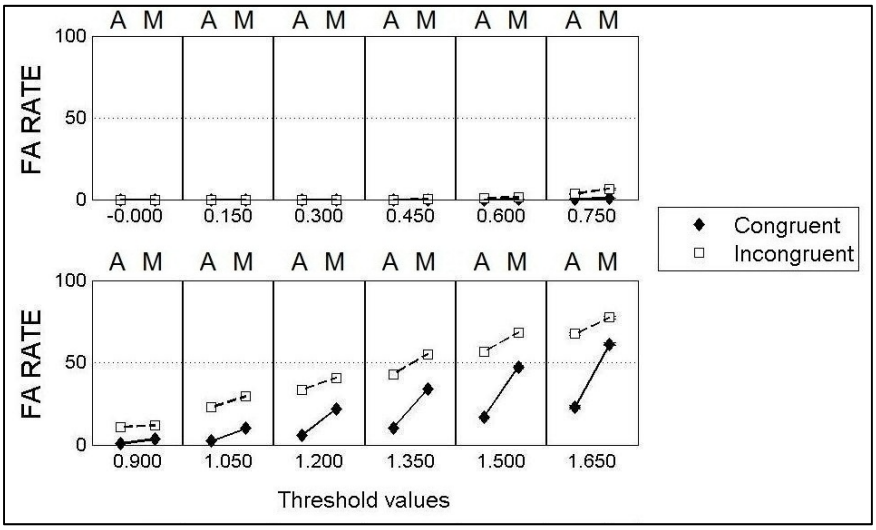
We have not yet considered the possibility that our model could in fact implement the "influence" account also. As we explicitly point out in Section 4.3, given that selective attention and decision-making are as much unsolved problems as face perception is, we just implemented the simplest methods that simulate these. The core of our model lies in the large, coarse templates. Without altering this, could alternative methods at the selective attention or decision-making stages implement the "influence" account?

The "influence" account states that when the composites are misaligned, the influence of the bottom halves on the perception of the top halves is reduced. Another way of thinking about this is that the relative contribution or weight of the bottom halves is reduced. Since the "reduction" account states that misalignment reduces responses, could this reduction somehow lead to reduced relative weight of the bottom halves?

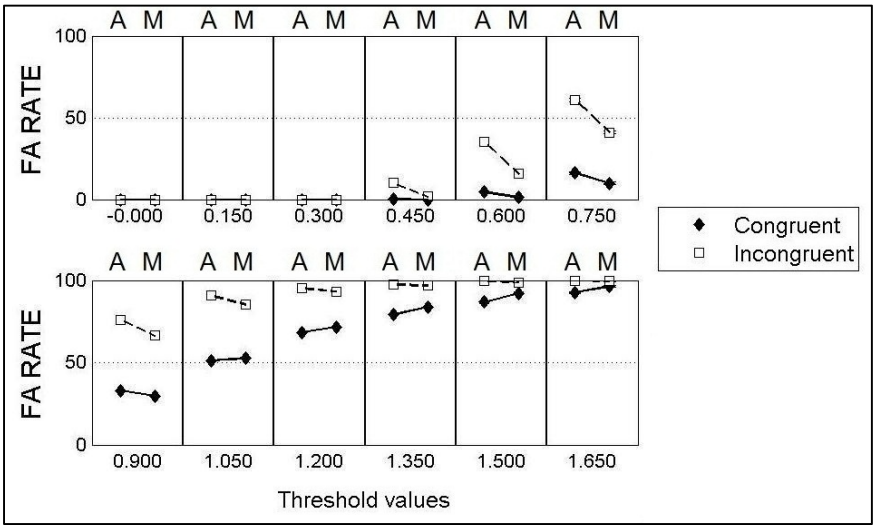
The normalized dot product (*ndp*) distance metric seems to implement this. (Note that the *ndp* measures similarity, so we define distance as 1 – *ndp*) A dot product is just a weighted linear sum. The *ndp* normalizes each vector (to have a norm of 1) first, before calculating the dot product. If the part of the vector corresponding to the response from the bottom half is reduced in magnitude, then normalization means that the other part (corresponding to the response from the top half) now has a larger relative contribution, similar to what happens according to the "influence" account. Note: unlike the Euclidean distance, the *ndp* distance generally need not decrease if the responses decrease, because of this normalization.



FS



LSF

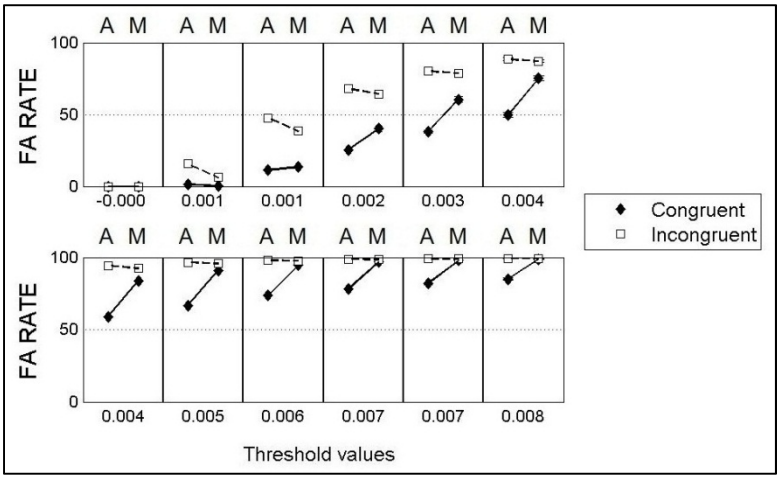


HSF

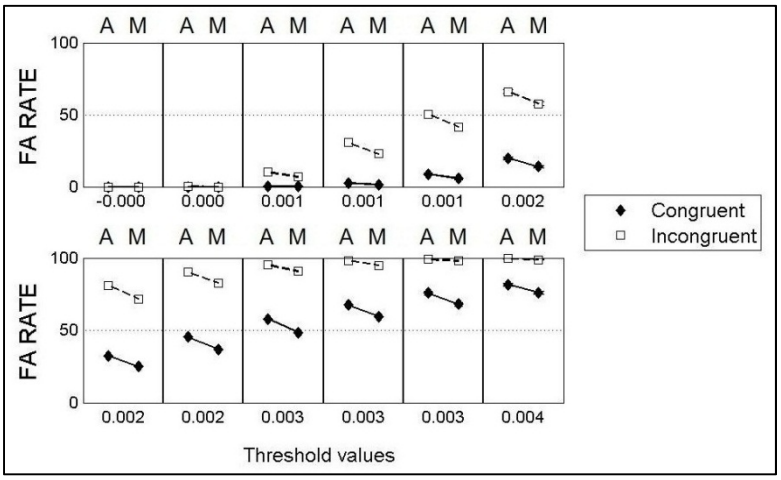
Figure 12.2. False alarm (FA) rates for FS (top), LSF (middle) and HSF (bottom) faces. A: aligned. M: misaligned.

Accordingly, we see from Fig. 12.3 (top) that when *ndp* is used as the distance metric, the incongruent FA rate does in fact decrease (especially at low FA rates), as per the "influence" account. (Compare this to Fig. 12.2 top for Euclidean distance). Crucially, even when using *ndp*, both the "partial" design misalignment effect and the "complete" design (congruency x alignment) interaction are still found (results not shown). In other words, the core of our model is consistent with both the "reduction" and "influence" accounts of holistic processing, and the only difference is in the distance metric used.

Interestingly, for small, fine features, while neither the misalignment effect nor the (congruency x alignment) interaction is found (results not shown), using *ndp* also produces an incongruent FA rate decrease (Fig. 12.3 bottom), similar to large, coarse features. This phenomenon needs further investigation, but it may be an interesting and counter-intuitive prediction of the model.



Large, coarse features



Small, fine features

Figure 12.3. False-alarm rates using the *ndp* distance metric for large, coarse features (top) and small, fine features (bottom). A: aligned. M: misaligned.

We have also not yet considered another question: is it necessarily the case that our model (even with the regular Euclidean distance metric) does not show "influence-like" behavior? From Fig. 12.2, we see that depending on the SF condition, our model (using large, coarse features nd Euclidean distance) can show behavior consistent with either "reduction" (FS and LSF) or "influence" (HSF) accounts. Amazingly, this is exactly what was found by Cheung et al. (2008), as shown in Fig. 12.4. We see that there is striking resemblance between Figs. 12.2 and 12.4. For any given threshold, not only does the FA rate generally rise from FS to LSF to HSF, the FA rates for congruent trials are also generally higher than for incongruent trials. In particular, again using thresholds of 0.600 and 0.750 (as we did in Chapter 9), we see that for the incongruent trials, there is little change in FA rates as a result of misalignment for FS and LSF. However, for HSF, the FA rates decrease, consistent with the "influence" account (p-value not reported in Cheung et al. 2008).

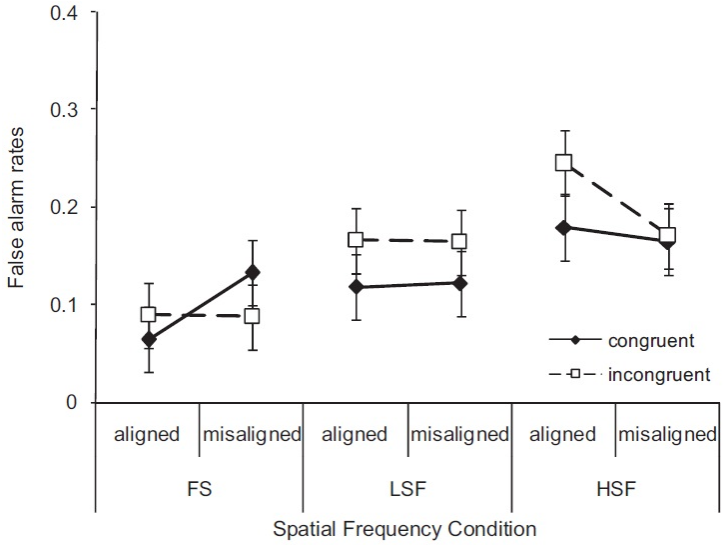


Figure 12.4. Empirical false-alarm rates for FS, LSF and HSF faces. Error bars show 95% confidence intervals of the 3 x 2 x 2 within-subjects interaction effect. Figure reproduced from Cheung et al. (2008). See p.16 for copyright notice.

12.6.3 Intuitive versus actual predictions

We seem to have contradicted our predictions in Section 12.4. If it were not actually clear that our model would make these predictions, or that it were actually different from the "influence" account, why did we use the "reduction" account to make predictions?

The "reduction" account was the intuitive way of explaining the behavior of our model, so that is easy to understand the misalignment effect and congruency x alignment interaction. However, an intuitive explanation is not a model. The point of having quantitative models is precisely because intuitive explanations are sometimes insufficient. Having quantitative models forces us to be explicit about assumptions and design decisions, such as which distance metric to use. Intuitive

explanations conveniently put these aside and ten to rely on human qualitative reasoning being adequate and infallible.

Thus, we wanted to illustrate the point pedagogically, that without a quantitative implementation, a qualitative model – such as the "influence" account – should not be seriously compared to a quantitative model. What might seem to be "predicted" by a qualitative model (e.g. the "reduction" account) may in reality be highly dependent on implicit assumptions that have not been thoroughly considered.

12.7 Proposed experiments

In Section 12.5, we said that the empirical evidence for some of the "reduction" account predictions was encouraging but far from conclusive. In light of the modeling results in Section 12.6, it may seem redundant to test the prediction for the incongruent-different trials. However, there is still a clear model prediction for the congruent-different trials that is robust to distance metric and is not found for small, fine features (see Figs. 12.2 top and 12.3). Note that if this prediction (increase in congruent FA rate for misaligned trials) is found empirically, it would not disprove the "influence" account per se (which makes no prediction), but would be evidence supporting the model (whether it is more "reduction"-like or "influence"-like is beside the point).

In theory, all the existing CFE studies (both "partial" and "complete) already have the raw data to test this prediction, even though only a handful report these results (i.e. the studies in Table 12.1). However, in practice, there is a problem that needs to be tackled. The accuracy in the "different" trials is often very good, possible leading to a ceiling effect (equivalently, floor effect for FA rate), as indicated in Table 12.1. This is also evident in the modeling results (Figs. 12.2 top and 12.3 top).

Therefore, a targeted study may need to intentionally increase FA rates, e.g. perhaps by using faces that differ subtly only in terms of second-order configural changes. However, this needs to be carefully calibrated to avoid overall chance-level performance. Apart from that, the guidelines suggested in Section 2.5.7 may help to isolate face-like intrinsic holistic processing.

12.8 Chapter summary

We compared the "reduction" account of the CFE with the more widespread "influence" account. We first show that the "reduction" account is better in terms of generalizability. Nonetheless, we proceed to compare the predictions from the two accounts. We find that there is some empirical support for the "reduction" account, but more targeted experiments are needed. We then proceed to show that the two accounts may actually differ in the "decision-making" stage, rather than the core "holistic processing" stage. Furthermore, we show that our model can actually show characteristics of both accounts under different circumstances. Overall, this chapter illustrates some pitfalls of relying solely on intuitive "mental models" to understand phenomena and make predictions.

Chapter 13: Discussion

Chapter abstract

In this final chapter, we begin by reiterating the main problem(s) that this thesis tries to address. We then propose a new theory of face processing based on our modeling work, and highlight the unique and novel contributions of our work. We proceed to discuss the broader implications of the theory and of our results, and then present a collection of predictions that can be tested empirically to validate our theory/model. Finally, we end off by listing several avenues for future work.

Chapter contents

THIS PAGE HAS BEEN

INTENTIONALLY LEFT BLANK

Chapter 13: Discussion

13.1 Linking the *what*, *how* and *why* of face processing

One of the biggest problems in the current understanding of face processing is the lack of a cohesive picture of how various facets of face processing relate to each other. For example, research on holistic/configural processing is notably disjoint from research on face-space/norm-coding (McKone 2009a). There is also little research relating face-space/norm-coding to image manipulations such as spatial frequency filtering and contrast reversal. Even within the holistic/configural ambit, the relationship between these two types of processing is still unclear (Maurer et al. 2002).

Furthermore, within each type of processing, we are far from achieving a cohesive understanding that links effect to process to cause (i.e. what, how and why). For example, Zhang & Cottrell (2005) suggest that holism arises because holistic features are good for identification, whereas Tsao & Livingstone (2008) suggest that holism arises from detection. For configural processing, it is unclear whether sensitivity to configural changes is due to the second-order properties of face stimuli, the task of identification, or some combination of the two.

This thesis is an initial attempt to address the key problem of the lack of an overarching framework to understand face processing. We started off by showing that a biologically plausible model of visual processing could replicate a key signature of holistic processing: the CFE Chapter 5). We then proceeded to show that the model could account for the relationship between holistic processing and several stimulus manipulations: inversion, contrast reversal, and spatial frequency filtering (Chapters 6 to 9). We also verified that the model can replicate the FIE (Chapter 10). Importantly, the model accounts for the difference between face-like and object-like processing for both the CFE and FIE. In Chapter 11, we then made preliminary attempts to address some of the issues mentioned earlier. We showed that our holistic model also performed configural processing. Furthermore, this configural processing shows characteristics of norm-based coding in face space. Finally, we showed that this holistic/configural/norm-coding model is related to both detection and identification of faces. In short, it appears that we may have found a model that is potentially capable of uniting all the different major aspects of face processing (less expression, gaze, etc.).

If this model does in fact truly account for all these aspects of face processing, what does that imply in terms of a theoretical understanding of face processing? We discuss this next.

13.2 A new theory of face processing

Here, we present a new theory regarding the mechanisms underlying face processing. This theory is, on the whole, not a radical departure from current theories, and has elements in common with some. Crucially, however, it is unprecedented in its ambition and scope, covering all the major aspects of identity-related face processing, i.e. excluding expression, gaze, etc. It is

also unique in its view of the effects of inversion, and of the key difference between face and object processing.

The heart of the theory is that what makes face processing different from object processing is the existence of large, coarse (LSF) templates for faces. Small, fine (HSF) templates exist for both faces and objects.

When any image is presented, all of these templates are used. Face images match all of these templates relatively well, possibly even the small, fine object templates. In contrast, non-face images only match the small, fine templates well (possibly including those for faces). Therefore, "face-specific processing" is really just the natural by-product of the fact that only faces match the large, coarse (face-tuned) templates well. In other words, what's special about face processing lies in the templates, not the template-matching process or some other kind of "special processing". The difference is quantitative rather than qualitative; there is likely to be a continuum between large and small templates in reality.

The large, coarse face templates act as a gateway leading to further face processing, such as gaze-tracking or expression recognition. This gating is implicit, and is simply the result of the neurons that correspond to these templates (henceforth "face cells") firing at best weakly in response to non-face stimuli. Importantly, the large, coarse templates support both face detection and identification, but they do not perform either task per se. These tasks are performed by downstream neurons (possibly in the prefrontal cortex) using the information from the face cells.

"Holistic processing" is simply the result of templates being large; the actual processing is just regular template-matching. "Holism" is not about literal wholes. "Holistic perception" arises from the fact that the individual contributions of local face regions cannot be titrated out based on the response of a face cell. Face processing is both "holistic" and "partistic", because face-related templates include both large, coarse ones and small, fine ones. Importantly, the distinction between large/coarse and small/fine is likely to be artificial; a continuum may exist.

"Configural processing" is simply the result of face cells being sensitive to second-order configural changes; the actual processing is, again, just regular template-matching. Configural and holistic effects are therefore just different phenomena arising from the same root cause – the use of large, coarse templates. The difference is in the stimulus changes; processing is identical.

Inversion does not really "disrupt" any processing. Inverted faces elicit reduced responses from face cells; the processing is, as always, just regular template-matching. Inverted faces are harder to discriminate because the reduced responses are either more similar to each other, are more susceptible to noise, or both. Objects exhibit smaller inversion effects because the responses of small, fine templates are reduced by a lesser amount. Inverted faces are processed "like objects" only to the extent that responses of small, fine templates become relatively more prominent.

The space in "face space" arises simply from the physical properties of face stimuli. "Face space" is manifested psychologically and behaviorally because large, coarse templates preserve the structure of that space. Norm-based coding relies on an implicit norm, not an explicit one.

More speculatively: large, coarse templates exist for faces due to a combination of innate factors, low visual acuity in infancy, ubiquity of faces, task demands and the physical properties of face stimuli (see Section 13.4.6 for longer discussion). The main locus of large, coarse face templates is the right FFA, which is biased towards LSFs. The small, fine face templates correspond to the left FFA, the OFA, or both.

13.3 Contributions

The main contribution of this thesis is that, for the very first time, we have demonstrated a model that may unify all the major aspects of face processing in a single framework.

Starting with holism, we show that our "large, coarse template" theory of face processing explains the difference between "face-like" and "object-like" processing for both the FIE and CFE. Importantly, because we kept the stimuli constant while changing the "processing style", we avoided the common confound of differences in physical stimulus characteristics. We also linked holism to the effects of inversion, spatial frequency filtering and contrast reversal.

Crucially, we linked the two main dominant frameworks: holistic/configural processing and face-space processing. With the same large, coarse templates used to account for holistic/configural processing, we showed that various aspects face-space/norm-coding can be reproduced also.

Another important contribution is the fact that our model links the properties of face-selective single-neurons to behavioral effects for face processing. Starting from the responses of individual model units, we provided mechanistic, step-by-step accounts of how these behavioral effects arose.

13.4 Implications

In this section, we elaborate on certain aspects of our theory that run counter to prevailing ideas, are more speculative, or are more broadly applicable beyond face processing.

13.4.1 The "single face" of configural processing

An influential review by Maurer et al. (2002) is titled "The many faces of configural processing". They distinguish between three kinds of "configural processing": first-order processing (related to detection), holistic processing, and second-order configural processing. Based on behavioral markers, interaction with stimulus manipulations, and developmental trajectories, Maurer et al. (2002) argue that these are distinct forms of processing.

In contrast, we propose that there is essentially only one form of processing (template matching), and these supposedly distinct forms of processing are simply different manifestations of the same process under different stimulus and task conditions.

The seemingly different developmental time-courses may just be incidental. We have not proposed a developmental aspect to our theory, but for now, let's suppose that the number of large, coarse templates increases with development. This increase in number of templates may have different effects on different tasks. Detection may not improve much, once a certain number of templates are present. On the other hand, identification (and sensitivity to configural changes) may require more templates before performance plateaus.

There are two caveats to our claim that there is just one form of processing. Firstly, the dynamics of the neural response may influence the information content available in face cells. Tsao et al. (2006) found that identification seems to have a longer latency than detection. We believe that this may simply be the result of gradually sharper tuning, rather than different processing. This sharpening of tuning could be a deliberate, general strategy not specific to faces, possibly linked to the notion of coarse-to-fine processing. Importantly, the process – template matching – is identical throughout.

The second caveat is that the relative contributions of small, coarse templates (both for faces and objects) may vary as a function of time-course, stimulus properties, task, and the responses to the large, coarse templates. This means that there may be apparent differences in "processing style" under different circumstances, but the differences are just quantitative. For example, if the behavioral task is fine discrimination, small, fine templates may be more heavily relied on, compared to when the behavioral task is rapid face/non-face detection.

13.4.2 Holism is not about wholes, and it is not all-or-none

According to our theory, "holism" is somewhat of a misnomer, because it is not really about wholes. Historically, "holistic processing" is simply a term used to describe the unobservable psychological construct that was created to explain certain results in behavioral experiments (Richler et al. 2011b). There is really no evidence at all to suggest that face processing mandatorily involves 100% of a given face stimuli, rather than say 90% or 80%.

It is unclear how literally the notion of "wholes" is generally taken, but the widespread usage of terms like "undecomposed", "unified", "unitary", "single", "all parts" and "as a whole" seems to suggest that this idea is taken seriously by some, if not many. This is rather mystifying, since faces can obviously be processed "part-istically" too. If faces were literally processed only as single unitary wholes, then we would only be able to perceive eyes and noses when faces were inverted (i.e. when "holism is disrupted").

Mechanistically, there are two versions of what it means for faces to be processed "as wholes". The first is that all parts of a face are processed together, and all non-face parts are excluded. One problem with this idea is that it does not exclude scrambled faces from being processed holistically. More importantly, it cannot account for the holistic processing of blurred faces, in which the face parts are only recognizable in the context of the face. This leads to the second version, in which face detection is required. However, faces are not an unambiguously defined category with sharp boundaries (Meng et al. 2012), and it would seem rather un-parsimonious to suggest a sharp switch in processing style occurs when some threshold of face-ness is crossed.

This brings us to another point, which is that according to our theory, holism is not all-or-none (nor is face processing in general; Bukach et al. 2006). Instead of having two distinct separate processing styles (likely performed by distinct neuronal populations), our theory simply proposes that the degree of holism is continuous. It is a function of both the face templates and the stimuli. Larger templates are more "holistic"; stimuli that match the templates less are less "holistically processed". While the templates can support detection, their responses are graded, not binary. This sidesteps the issue of whether line drawings, cartoon depictions, partial faces (and an infinite number of potentially face-like stimuli) should be considered faces or not. It is simply a matter of how well they match the face templates, so the issue is merely a quantitative one.

13.4.3 Link between discriminability and neural response

Beyond face processing, our theory has some interesting implications for visual recognition in general. For faces, since the templates are tuned to upright faces, they respond to inverted faces less strongly. Therefore, when inverted faces have to be discriminated, the elicited patterns of responses are more similar to each other, and are less discriminable. For the small, fine templates, inversion has less of a detrimental effect (e.g. since an inverted eye still somewhat resembles an eye). These findings are not profound or novel, and have been demonstrated before (Zhang & Cottrell 2004, 2006, Jiang et al. 2006).

What's interesting is the link to other phenomena. Take the other-race effect (ORE), for example. Face templates may be particularly attuned to own-race faces; therefore other-race faces may elicit smaller responses, which lead to less discriminability.

Another interesting link is that to expertise. One explanation of why discrimination (or identification) performance improves with training is that the classification boundaries become appropriately adjusted through supervised training. However, another explanation (not mutually exclusive) is that the templates are adjusted in an unsupervised manner. Initially, generic templates may not respond strongly to the stimuli in question. However, through exposure, these templates become tuned to the stimuli, and therefore respond more strongly (Sigala & Logothetis 2002). This leads to better discrimination.

13.4.4 The units of perception and attention

While our work on holism models purely the behavioral aspects (e.g. reproducing the CFE), there are interesting implications for the more subjective, perceptual aspects. The CFE is not simply an effect that arises from subjects making same-different discriminations; it arises from the experience of the face halves forming a combined percept. This does not happen when the face halves are inverted. The "perceptual field" hypothesis (Rossion 2009) reflects precisely these findings.

However, what exactly is a "perceptual field"? How does it come about? Why is it smaller for inverted faces and objects? Our theory proposes a very simple and straightforward answer. Perception must arise from the activity of some population of neurons. The "basic unit" of perception therefore corresponds to the activity of one neuron. If the neurons are the "face cells"

that implement large, coarse template matching, then each neuron's template is a largish face region. Putting these ideas together, the basic, indivisible unit of perception corresponds to a largish face region! In other words, the halves of a composite face are perceptually combined because each unit of perception covers a large portion of the face, and many units will span both halves. (Note that the small, fine templates contribute to perception too, but we ignored them for simplicity)

For inverted faces, because the small, fine templates are less affected by inversion, their responses are now larger than those of large, coarse templates. Perception is therefore dominated by these responses. Accordingly, the "perceptual field" is thus, overall, smaller than for upright faces because of the template sizes. The same idea applies for perception of objects.

A highly related topic is that of attention. Like for perception, the basic unit of attention must be the neuron. Using similar logic as for the case of perception, we argue that the "attentional field" for upright faces is larger than that for inverted faces and objects.

In our simulations, we chose to model attentional modulation at the pixel level, for reasons of simplicity and making minimal assumptions. In reality, we believe that the notion of a large "attentional field" for upright faces is precisely why subjects display the CFE despite explicit instructions to attend to only the top halves. Because many face cells are tuned to face regions that span both halves, the attentional system has no choice but to (at most) minimally modulate the responses of these face cells. The exact algorithm for attentional modulation is not important for explaining the CFE. What is important is that in comparison to inverted faces and objects, the (large, coarse) templates for upright faces span both halves in greater proportion and to a greater extent.

13.4.5 Faces, faces, everywhere

One interesting phenomenon is that people seem to be able to perceive face-ness is many situations where there is clearly not a real (biological) face. While this phenomenon is clearly not solely restricted to faces (e.g. people can imagine many things from looking at clouds), it seems to be much more prevalent for faces.

Our theory posits that this is because of the coarseness of the face templates. While holism stems from the largeness rather than the coarseness of the templates, we argue for various reasons (see next section) that these templates should also be coarse. What this means is that many images will tend to activate these templates relatively strongly, as long as the images have the first-order configuration of faces. Importantly, this is not face detection per se, because it is clear that these face-like stimuli are not real faces. Interestingly, face-specific properties such as emotional expression can be perceived from such face-like images, further arguing against the notion that face detection acts as a binary gating mechanism for subsequent face-specific processing.

13.4.6 Why large, coarse templates?

This thesis was focused on the "how" (mechanisms) and the "what" (behavior, electrophysiology) of face processing. Here, we speculate about the "why" – specifically, why

would face templates be both large and coarse? We believe that it is a combination of stimulus properties, innate genetics, ubiquity, social demands, poor infant visual acuity, and task demands. The relative importance of these factors is debatable, but they all play a part.

The most fundamental factor is stimulus properties – the fact that faces have a common first-order configuration. If not for this fact, the other factors would most likely not come into play. Both largeness and coarseness are viable characteristics because of the common first-order configuration. For other stimulus classes that lack a first-order configuration (e.g. fruits, lamps, and chairs), large templates or coarse templates would be poorly informative.

There is evidence of innate genetic influences on face perception (Polk et al. 2007, Sugita 2008, McKone & Palermo 2010, Wilmer et al. 2010), and in particular, evidence for large, coarse templates (Valenza et al. 1996, Turati et al. 2002). This is linked to other factors such as ubiquity, social demands, stimulus properties and infant visual acuity. It would appear that there may be some evolutionary advantage to having good face recognition abilities, since faces are ubiquitous and their recognition is essential for social survival. However, if faces did not have a common first-order configuration, innate coding for coarse templates may not be viable. Moreover, due to poor visual acuity in infancy, it would be uneconomical – perhaps even detrimental – for innate face templates to be finely coded.

Even if there were no innate specification of face templates (and face recognition abilities were developed purely through interaction with the environment), the ubiquity of faces during infancy would still seem to dictate that face templates would arise early in development. They would be coarse templates, as a result of poor visual acuity. Furthermore, since small, coarse templates would be rather less informative than large, coarse ones, the latter would be preferred. Interestingly, it has been found that subjects with early visual deprivation do not show the CFE misalignment effect (Le Grand et al. 2004). Later in development, more non-face objects are viewed and need to be recognized – at which time visual acuity may be sufficient to support small, fine templates (for both faces and non-faces).

Finally, a combination of task demands and stimulus properties may also necessitate large, coarse templates. More so than for any other stimulus class, faces commonly need to be identified, not just detected. At the same time, individual faces vary strongly (though not solely) in second-order configuration. Discrimination of configural differences may rely strongly on large, coarse templates.

13.5 Predictions

Models – whether qualitative mental models or quantitative computational models – are important as tools to help put together a coherent understanding. Equally important is the role of models in suggesting experiments and making predictions to help advance that understanding. In this section, we attempt to do precisely that.

Our predictions come in various degrees of specificity. Also, not all stem directly from the quantitative model that was implemented; some are from the more qualitative theoretical

framework proposed earlier in this chapter. The more speculative or vague predictions are indicated by an asterisk (*).

13.5.1 General predictions

The CFE is a differential effect. While some evidence of this already exists (see Chapter 2), we believe that stronger evidence can be found by using unambiguously non-face stimuli that vary in identity solely in terms of second-order configuration. Examples could include dot patterns (e.g. like the stimuli used in Farah, Tanaka & Drain 1995)

Holism for face-like stimuli. Stimuli that look like faces, but are clearly not (e.g. eyes replaced with other objects, Donnelly et al. 1994) should also elicit a FIE and CFE, but probably of a smaller magnitude. This is because the large, coarse templates are activated relatively strongly. In order to maximize this effect, the task should be discrimination of configural changes, and conducted should be under feedforward conditions (e.g. short presentation times, masking).

Measures of holism correlate with face-ness ratings. Related to the previous prediction, subjects' rating of face-ness (of noise-masked or LSF stimuli, for example) should correlate with measures of holism on a trial-by-trial basis. For example, for a single trial using two stimuli that rated as highly face-like, subjects should (on average) exhibit the CFE. This is unlike the current paradigms, which average over trials that unambiguous faces.

Behavioral CFE magnitude correlates with neural reduction on a trial-by-trial basis. Our account of the CFE banks on reduction of responses to inverted and misaligned faces. Measures of neural activity (e.g. BOLD, N170, spike count) in "holistic" brain areas (e.g. FFA, MF/ML) should reflect this reduction on individual trials that show the CFE.

Measures of holism are correlated with "largeness". Neurons or brain areas are usually characterized as responsive to parts or wholes in a binary manner. More sensitive, continuous characterization in terms of "largeness" (e.g. size of optimal face stimulus) should correlate with measures of holism, since holism is not all-or-none.

Face-space adaptation to contrast-reversed and regular faces is similar. Responses to inverted and upright faces are negatively correlated, which is why adaptation effects appear to be distinct for inverted and upright faces. However, responses to contrast-reversed and regular faces are positively correlated. Therefore, neurons that are strongly activated by (and adapt to) regular faces, are also the ones that would have been strongly activated by contrast-reversed faces.

* Attentional "resilience" for faces. Since face cells are tuned to large face regions, faces may be more "resilient" than objects in terms of attentional modulation. Some examples might involve overlapping translucent stimuli, or scenarios involving attentional capture. However, it may be difficult to disentangle any results from the fact that faces are more salient due to social factors.

* The size of the face "perceptual field" is not retinotopic. Earlier, we stated that the perceptual field is larger for upright than inverted faces. However, since we believe that face templates are

scale-tolerant and not tied to retinotopic factors, these sizes are in relative terms (e.g. percentage of a whole face), not absolute ones (e.g. degree of visual angle).

13.5.2 Electrophysiology

The optimal stimuli for MF/ML are large faces regions. Freiwald et al. (2009) have shown that MF/ML face cells respond to at most four semantic face parts. We predict that by using a "growing" search paradigm (increasing the proportion of a face until responses plateau or decrease), the optimal stimulus should similarly be large (but not necessarily whole).

MF/ML neurons fire more strongly to aligned than misaligned composites. Our account of the CFE banks on reduction of single-neuron responses due to misalignment. Since MF/ML neurons seem to share many common characteristics as our large, coarse units, we predict that these neurons respond to misalignment like our units do, and are the neural basis of the CFE.

Neurons that are most strongly activated by upright faces are the ones most affected by inversion. According to our simulations (see Fig. 10.1), the responses to upright and inverted faces are negatively correlated for large, coarse templates.

13.5.3 fMRI

Voxels that span both halves will show the misalignment effect, but not others. Our account of the misalignment effect essentially arises from the fact that our large, coarse units span both halves of the composites. Regardless of specific brain region (OFA, FFA, fSTS), we predict that only individual voxels that respond significantly stronger to whole faces than either half alone will show the CFE. Furthermore, this is also true for voxels that respond significantly stronger to the middle region than top or bottom regions (equalized for size).

13.5.4 Behavior

Bias towards "same" for inverted faces. Cheung et al. (2008) found that for both "partial" and "complete" designs, misalignment shifted the bias towards "same". Our simulations predict that this shift should also happen for inversion.

13.6 Future work

13.6.1 Detection versus identification

While we found that large, coarse templates seem to be able to support both detection and identification, we have yet not actually trained classifiers to do so. We predict that performance on both tasks will be above chance (though not necessarily equally good). An interesting extension to this will be to examine performance as a function of tuning width. One possibility is that in the brain, initial broad tuning optimizes detection performance, while a gradual sharpening of tuning happens in order to better support identification.

13.6.2 Face space, norm-based coding and caricatures

In chapter 11, our adaptation results were not ideal in the sense that adaptation to both -3 and +3 face were in the same direction, and we hypothesized that this was due to the small and non-representative sample of faces used during template extraction. We should verify that a large, naturalistic set of faces are indeed normally distributed (as presumed), and then examine the templates extracted from this set of faces to see if these are also normally distributed in some manner.

We have not demonstrated the caricature effect. Modifications to the model, such as allowing the templates to "repel" each other through competitive interactions (Brunelli & Poggio 1993), may be required.

13.6.3 Featural versus configural processing

While similar models such as those of Zhang & Cottrell (2004, 2006) and Jiang et al. (2006) have replicated the behavioral findings regarding featural versus configural changes, our model has not. This will be an important sanity check for our model.

13.6.4 Other-Race Effect (ORE)

We can attempt to reproduce the ORE by extracting templates from faces of only one race, and then compare the discrimination performance on "own-race" versus "other-race" faces.

13.6.5 Computer Vision

Variants of our model have been used for object recognition tasks on standard Computer Vision (CV) datasets and achieved competitive performance (Serre et al. 2007, Mutch & Lowe 2008). If our model does indeed mimic the human visual system, then we would hope that it also performs well on standard CV datasets for faces as well. Unlike the approach taken in this thesis, substantial parameter tuning may be required. Importantly, both large/coarse and small/fine templates are likely to be needed for good performance. Interestingly, Pinto and colleagues (Pinto et al. 2011) have already achieved good face recognition performance using similar models. We have not yet compared the properties (e.g. template size) of our model and theirs.

13.7 Conclusion

In this thesis, we have concentrated on accounting for face processing using large, coarse features. However, at every step along the way, we have used small, fine features as a control, calling it "object-like" processing. While this thesis has already set the ambitious goal of providing a unified account of face processing, there is the remote possibility that if our model is correct in that face and object processing really just differ in the use of large, coarse templates versus small, fine templates, then perhaps we have also made the first step towards a unified account of face and object processing!

References

Amit, Y., & Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Research*, *43*(19), 2073-2088.

Anstis, S. (2005). Last but not least. *Perception*, *34*(2), 237-40.

Biederman, I, & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *352*(1358), 1203-19. The Royal Society.

Bruce, V, & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77 (Pt 3)*, 305-27.

Bruce, V, Burton, A. M., & Dench, N. (1994). What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *47*(1), 119-41.

Bruce, Vicki, Burton, A. M., & Hancock, P. J. (1995). Missing dimensions of distinctiveness. In Tim Valentine (Ed.), *Cognitive and Computational Aspects of Face Recognition: Explorations in Face Space* (pp. 138-158). London: Routledge.

Brunelli, R., & Poggio, T. (1993). Caricatural effects in automated face perception. *Biological Cybernetics*, *69*(3), 235-241. Springer Berlin / Heidelberg.

Bukach, C. M., Bub, D. N., Gauthier, I., & Tarr, M. J. (2006). Perceptual expertise effects are not all or none: spatially limited perceptual expertise for faces in a case of prosopagnosia. *Journal of Cognitive Neuroscience*, *18*(1), 48-63.

Bukach, C. M., Phillips, W. S., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Perception & Psychophysics*, *72*(7), 1865-74.

Calder, A. J., & Jansen, J. (2005). Configural coding of facial expressions: The impact of inversion and photographic negative. *Visual Cognition*, *12*(3), 495-518.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641-51.

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*(5163), 1333-6.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, *17*(21), 8621-44.

Carey, S., & Diamond, R. (1994). Are faces perceived as configurations more by adults than by children? *Visual Cognition*, *1*(2), 253-274. Psychology Press.

Carrasco, M, Penpeci-Talgar, C., & Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Research*, *40*(10-12), 1203-15.

Carrasco, Marisa, Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, *7*(3), 308-13.

Cheung, O. S., & Gauthier, I. (2010). Selective interference on the holistic processing of faces in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 448-61.

Cheung, O. S., Richler, J. J., Palmeri, T. J., & Gauthier, I. (2008). Revisiting the role of spatial frequencies in the holistic processing of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1327-36.

Cheung, O. S., Richler, J. J., Phillips, W. S., & Gauthier, I. (2011). Does temporal integration of face parts reflect holistic processing? *Psychonomic Bulletin & Review*, *18*(3), 476-83.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & Psychophysics*, *58*(4), 602-12.

Cottrell, G. W., Branson, K. M., & Calder, A. J. (2002). Do Expression and Identity Need Separate Representations? *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 238 - 243).

Dahl, C. D., Logothetis, N. K., & Hoffman, K. L. (2007). Individuation and holistic processing of faces in rhesus monkeys. *Proceedings. Biological Sciences / The Royal Society*, *274*(1622), 2069-76.

Dailey, M. N., & Cottrell, G. W. (1998). Task and Spatial Frequency Effects on Face Specialization. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference* (Vol. 8, pp. 17-23). MIT Press.

Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, *12*(7-8), 1053-1074.

Dakin, S. C., & Watt, R. J. (2009). Biological "bar codes" in human faces. *Journal of Vision*, *9*(4), 2.1-10.

de Heering, A., & Rossion, B. (2008). Prolonged visual experience in adulthood modulates holistic face perception. *PLoS One*, *3*(5), e2317.

de Heering, A., Houthuys, S., & Rossion, B. (2007). Holistic face processing is mature at 4 years of age: evidence from the composite face effect. *Journal of Experimental Child Psychology*, *96*(1), 57-70.
de Heering, A., Rossion, B., Turati, C., & Simion, F. (2008). Holistic face processing can be independent of gaze behaviour: evidence from the composite face illusion. *Journal of Neuropsychology*, *2*(Pt 1), 183-95.
De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545-59.
Deruelle, C., Rondan, C., Gepner, B., & Tardif, C. (2004). Spatial frequency and face processing in children with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *34*(2), 199-210.
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), 2051-62.
Donnelly, N., Humphreys, G. W., & Sawyer, J. (1994). Stimulus factors affecting the categorisation of faces and scrambled faces. *Acta Psychologica*, *85*(3), 219-34.
Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 628-34.
Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482-98.
Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(47), 17245-50.
Fiorentini, A., Maffei, L., & Sandini, G. (1983). The role of high spatial frequencies in face perception. *Perception*, *12*(2), 195-201.
Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: direct evidence. *Perception*, *29*(2), 159-70.
Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187-96.
Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193-202.
Galper, R. E. (1970). Recognition of faces in photographic negative. *Psychonomic Science*, *19*(4), 207-208.
Gao, Z., Flevaris, A. V., Robertson, L. C., & Bentin, S. (2011). Priming global and local processing of composite faces: revisiting the processing-bias effect on face perception. *Attention, Perception & Psychophysics*, *73*(5), 1477-86.
Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, *103*(2), 322-30.
Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 431-46.
Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training "greeble" experts: a framework for studying expert object recognition processes. *Vision Research*, *38*(15-16), 2401-28.
Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, *6*(4), 428-32.
Gauthier, I., Klaiman, C., & Schultz, R. T. (2009). Face composite effects reveal abnormal face processing in Autism spectrum disorders. *Vision Research*, *49*(4), 470-8.
Gauthier, I., Tarr, M. J., & Bub, D. (2010). *Perceptual Expertise: Bridging Brain and Behavior*. Oxford University Press.
George, N., Dolan, R. J., Fink, G. R., Baylis, G. C., Russell, C., & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nature Neuroscience*, *2*(6), 574-80.
Gilad, S., Meng, M., & Sinha, P. (2009). Role of ordinal contrast relationships in face encoding. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(13), 5353-8.
Goffaux, V. (2009). Spatial interactions in upright and inverted faces: re-exploration of spatial scale influence. *Vision Research*, *49*(7), 774-81.
Goffaux, V., & Dakin, S. C. (2010). Horizontal information drives the behavioral signatures of face processing. *Frontiers in Psychology*, *1*, 143. Frontiers Media SA.
Goffaux, V., & Rossion, B. (2006). Faces are "spatial"--holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 1023-39.
Goffaux, V., & Rossion, B. (2007). Face inversion disproportionately impairs the perception of vertical but not horizontal relations between features. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 995-1002.

Goffaux, V., Hault, B., Michel, C., Vuong, Q. C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception, 34*(1), 77-86.
Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research, 39*(21), 3537-60.
Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology, 35*(1), 96-111.
Hancock, P J, Burton, A. M., & Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory & Cognition, 24*(1), 21-40.
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), 189-99.
Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233.
Hayes, A. (1988). Identification of two-tone images; some implications for high- and low-spatial-frequency processes in human vision. *Perception, 17*(4), 429-36.
Hayes, T., Morrone, M. C., & Burr, D. C. (1986). Recognition of positive and negative bandpass-filtered images. *Perception, 15*(5), 595-602.
Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*(2), 181-97.
Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23*(1), 65-74.
Hole, G. J., George, P. A., & Dunsmore, V. (1999). Evidence for holistic processing of faces viewed as photographic negatives. *Perception, 28*(3), 341-59.
Hsiao, J. H., & Cottrell, G. W. (2009). Not all visual expertise is holistic, but it may be leftist: the case of Chinese character recognition. *Psychological Science, 20*(4), 455-63.
Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*(5749), 863-6.
Itier, R. J., & Taylor, M. J. (2002). Inversion and contrast polarity reversal affect both encoding and recognition processes of unfamiliar faces: a repetition study using ERPs. *NeuroImage, 15*(2), 353-72.
Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using FMRI and behavioral techniques. *Neuron, 50*(1), 159-72.
Kanwisher, N., & Yovel, G. (2009). Face Perception. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition, 68*(1), B1-11.
Kemp, R., McManus, C., & Pigott, T. (1990). Sensitivity to the displacement of facial features in negative and inverted images. *Perception, 19*(4), 531-43.
Kemp, R., Pike, G., White, P., & Musselman, A. (1996). Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues. *Perception, 25*(1), 37-52.
Khurana, B., Carter, R. M., Watanabe, K., & Nijhawan, R. (2006). Flash-lag chimeras: the role of perceived alignment in the composite face effect. *Vision Research, 46*(17), 2757-72.
Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*(3), 856-867.
Le Grand, R, Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Neuroperception. Early visual experience and face processing. *Nature, 410*(6831), 890.
Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science, 15*(11), 762-8.
Leonard, H. C., Annaz, D., Karmiloff-Smith, A., & Johnson, M. H. (2011). Developing spatial frequency biases for face recognition in autism and Williams syndrome. *Journal of Autism and Developmental Disorders, 41*(7), 968-73.
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), 89-94.
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), 572-5.
Lewis, M. B., & Johnston, R. A. (1999). A unified account of the effects of caricaturing faces. *Visual Cognition, 6*(1), 1-41. Psychology Press.
Liu, C. H., & Chaudhuri, A. (1997). Face recognition with multi-tone and two-tone photographic negatives. *Perception, 26*(10), 1289-96.

Logothetis, N K, & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, *5*(3), 270-88.
Logothetis, N K, Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552-63.
Macchi Cassia, V., Picozzi, M., Kuefner, D., Bricolo, E., & Turati, C. (2009). Holistic processing for faces and cars in preschool-aged children and adults: evidence from the composite effect. *Developmental Science*, *12*(2), 236-48.
McKone, E. (2008). Configural processing and face viewpoint. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 310-27.
McKone, E. (2009). Holistic processing for faces operates over a wide range of sizes but is strongest at identification rather than conversational distances. *Vision Research*, *49*(2), 268-83.
McKone, E. (2010). Integrating holistic processing and face-space approaches to the coding of facial identity. *Journal of Vision*, *9*(8), 539-539.
McKone, E., & Palermo, R. (2010). A strong role for nature in face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 4795-6.
McKone, E., & Robbins, R. (2007). The evidence rejects the expertise hypothesis: reply to Gauthier & Bukach. *Cognition*, *103*(2), 331-6.
McKone, E., & Robbins, R. (2011). Are Faces Special? In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *Oxford Handbook of Face Perception* (pp. 149-176). Oxford University Press.
McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin & Review*, *16*(5), 778-97.
McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8-15.
Mel, B. W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, *9*(4), 777-804.
Meng, M., Cherian, T., Singal, G., & Sinha, P. (2012). Lateralization of face processing in the human brain. *Proceedings. Biological Sciences / The Royal Society*.
Michel, C., Rossion, B., Han, J., Chung, C.-S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608-15.
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), 1355-9.
Mondloch, C. J., & Maurer, D. (2008). The effect of face orientation on holistic processing. *Perception*, *37*(8), 1175-86.
Mutch, J., & Lowe, D. G. (2006). Multiclass Object Recognition Using Sparse, Localized Features. *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 11-18). New York.
Mutch, J., & Lowe, D. G. (2008). Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields. *International Journal of Computer Vision*, *80*(1), 45-57. Springer Netherlands.
Mutch, J., Knoblich, U., & Poggio, T. (2010). CNS: a GPU-based framework for simulating cortically-organized networks. CBCL Memo 286. CSAIL Memo 2010-013. MIT. Cambridge, MA.
Nasanen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, *39*(23), 3824-33.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353-383.
Nederhouser, M., Yue, X., Mangini, M. C., & Biederman, I. (2007). The deleterious effect of contrast reversal on recognition is unique to faces, not objects. *Vision Research*, *47*(16), 2134-42.
Ohayon, S., Tsao, D., & Freiwald, W. (2010). Contrast tuning in face cells - Evidence for region based encoding. *Society for Neuroscience Annual Meeting*.
Oliva, A, & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72-107.
Perrett, D. I., & Oram, M. W. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, *11*(6), 317-333.
Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, *47*(3), 329-42.

Perrett, D. I., Mistlin, A. J., Chitty, A. J., Smith, P. A., 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), 245-58.

Pestilli, F., & Carrasco, M. (2005). Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision Research*, *45*(14), 1867-75.

Pinto, N., Stone, Z., Zickler, T., & Cox, D. (2011). Scaling Up Biologically-Inspired Computer Vision: A Case Study in Unconstrained Face Recognition on Facebook. *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 25-32).

Polk, T. A., Park, J., Smith, M. R., & Park, D. C. (2007). Nature versus nurture in ventral visual cortex: a functional magnetic resonance imaging study of twins. *Journal of Neuroscience*, *27*(51), 13921-5.

Ramon, M., & Rossion, B. (2012). Hemisphere-dependent holistic processing of familiar faces. *Brain and Cognition*, *78*(1), 7-13.

Ranzato, M. A., Huang, F., Boureau, Y.-L., & LeCun, Y. (2007). Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. *Computer Vision and Pattern Recognition, IEEE Conference on*.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977-87.

Rhodes, G., Jeffery, L., Watson, T. L., Jaquet, E., Winkler, C., & Clifford, C. W. G. (2004). Orientation-contingent face aftereffects and implications for face-coding mechanisms. *Current Biology*, *14*(23), 2119-23.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2007). Distinguishing Between Perceptual and Decisional Sources of Holism in Face Processing. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Richler, J. J., Gauthier, I., Wenger, M. J., & Palmeri, T. J. (2008a). Holistic processing of faces: perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 328-42.

Richler, J. J., Tanaka, J. W., Brown, D. D., & Gauthier, I. (2008b). Why does selective attention to parts fail in face processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1356-68.

Richler, J. J., Bukach, C. M., & Gauthier, I. (2009a). Context influences holistic processing of nonface objects in the composite task. *Attention, Perception & Psychophysics*, *71*(3), 530-40.

Richler, J. J., Cheung, O. S., Wong, A. C.-N., & Gauthier, I. (2009b). Does response interference contribute to face composite effects? *Psychonomic Bulletin & Review*, *16*(2), 258-63.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009c). Holistic processing of faces happens at a glance. *Vision Research*, *49*(23), 2856-61.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011a). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464-71.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011b). Beliefs alter holistic face processing ... if response bias is not taken into account. *Journal of Vision*, *11*(13), 17.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011c). Inverted faces are (eventually) processed holistically. *Vision Research*, *51*(3), 333-42.

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011d). Perceptual Expertise as a Shift from Strategic Interference to Automatic Holistic Processing. *Current Directions in Psychological Science*, *20*(2), 129-134.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019-25.

Riesenhuber, M., & Poggio, T. (2003). How the Visual Cortex Recognizes Objects: The Tale of the Standard Model. In L. M. Chalupa & J. S. Werner (Eds.), *The Visual Neurosciences* (pp. 1640-1653). Bradford Books.

Riesenhuber, M., Jarudi, I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proceedings. Biological Sciences / The Royal Society*, *271 Suppl. 6*, S448-50.

Robbins, R., & McKone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition*, *88*(1), 79-107.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*(1), 34-79.

Rolls, E T, & Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research*, *65*(1), 38-48.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica*, *128*(2), 274-89.

Rossion, B. (2009). Distinguishing the cause and consequence of face inversion: the perceptual field hypothesis. *Acta Psychologica*, *132*(3), 300-12.

Rossion, B., & Boremanse, A. (2008). Nonlinear relationship between holistic processing of individual faces and picture-plane rotation: evidence from the face composite illusion. *Journal of Vision, 8*(4), 3.1-13.
Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 63-75.
Ruiz-Soler, M., & Beltran, F. S. (2006). Face perception: an integrative review of the role of spatial frequencies. *Psychological Research, 70*(4), 273-92.
Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Psychological and biological models.* MIT Press.
Russell, R., Sinha, P., Biederman, I., & Nederhouser, M. (2006). Is pigmentation important for face recognition? Evidence from contrast negation. *Perception, 35*(6), 749-59.
Russell, R., Biederman, I., Nederhouser, M., & Sinha, P. (2007). The utility of surface reflectance for the recognition of upright and inverted faces. *Vision Research, 47*(2), 157-65.
Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience, 30*(39), 12978-95.
Schiltz, C., & Rossion, B. (2006). Faces are represented holistically in the human occipito-temporal cortex. *NeuroImage, 32*(3), 1385-94.
Schiltz, C., Dricot, L., Goebel, R., & Rossion, B. (2010). Holistic perception of individual faces in the right middle fusiform gyrus as evidenced by the composite face illusion. *Journal of Vision, 10*(2), 25.1-16.
Schwaninger, A., Lobmaier, J. S., Wallraven, C., & Collishaw, S. (2009). Two routes to face perception: evidence from psychophysics and computational modeling. *Cognitive Science, 33*(8), 1413-40.
Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition, 69*(3), 243-65.
Serre, T., & Riesenhuber, M. (2004). Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model, and Implications for Invariant Object Recognition in Cortex. CBCL Memo 239. MIT. Cambridge, MA.
Serre, T., Oliva, A., & Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6424-9.
Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007b). Robust Object Recognition with Cortex-Like Mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29*(3), 411-426.
Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1977). *Perceiving and remembering faces.*
Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*(6869), 318-20.
Singer, J. M., & Sheinberg, D. L. (2006). Holistic processing unites face parts across time. *Vision Research, 46*(11), 1838-47.
Sinha, P. (2002). Qualitative Representations for Recognition. In H. H. Bülthoff, C. Wallraven, S.-W. Lee, & T. A. Poggio (Eds.), *Biologically Motivated Computer Vision* (Vol. 2525, pp. 129-146). Springer, Berlin Heidelberg.
Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE, 94*(11), 1948 - 1962.
Skottun, B. C., De Valois, R. L., Grosof, D. H., Movshon, J. A., Albrecht, D. G., & Bonds, A. B. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research, 31*(7-8), 1079-86.
Soria Bauser, D. A., Suchan, B., & Daum, I. (2011). Differences between perception of human faces and body shapes: evidence from the composite illusion. *Vision Research, 51*(1), 195-202.
Subramaniam, & Biederman, I. (1997). Does contrast reversal affect object identification. *Investigative Opthalmology and Visual Science.*
Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences of the United States of America, 105*(1), 394-8.
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), 300-14.
Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109-39.
Tanaka, J. W, & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition, 25*(5), 583-92.
Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66*(1), 170-89.
Taubert, J., & Alais, D. (2009). The composite illusion requires composite face stimuli to be biologically plausible. *Vision Research, 49*(14), 1877-85.
Taubert, J., & Alais, D. (2011). Identity aftereffects, but not composite effects, are contingent on contrast polarity. *Perception, 40*(4), 422-36.

Tsao, D. Y., & Freiwald, W. A. (2006). What's so special about the average face? *Trends in Cognitive Sciences*, *10*(9), 391-3.
Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, *31*, 411-37.
Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670-4.
Tsao, D. Y., Cadieu, C., & Livingstone, M. S. (2010). Object Recognition: Physiological and Computational Insights. In Platt & Ghazanfar (Eds.), *Primate Neuroethology* (Vol. 8, p. 670). Oxford University Press.
Turati, C., Simion, F., Milani, I., & Umiltà, C. (2002). Newborns' preference for faces: what is crucial? *Developmental Psychology*, *38*(6), 875-82.
Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71-86.
Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682-7.
Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79 (Pt 4)*, 471-91.
Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *43*(2), 161-204.
Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 892-903.
Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience*, *6*(6), 624-31.
Vuong, Q. C., Peissig, J. J., Harrison, M. C., & Tarr, M. J. (2005). The role of surface pigmentation for recognition revealed by contrast reversal in faces and Greebles. *Vision Research*, *45*(10), 1213-23.
Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167-94.
Wallraven, C., Schwaninger, A., & Bülthoff, H. H. (2005). Learning from humans: computational modeling of face recognition. *Network: Computation in Neural Systems*, *16*(4), 401-18.
Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-77.
Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, *6*(4), 647-53.
Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 5238-41.
Wong, Y. K., & Gauthier, I. (2010). Holistic processing of musical notation: Dissociating failures of selective attention in experts and novices. *Cognitive, Affective & Behavioral Neuroscience*, *10*(4), 541-51.
Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009a). Conditions for facelike expertise with objects: becoming a Ziggerin expert--but which type? *Psychological Science*, *20*(9), 1108-17.
Wong, A. C.-N., Palmeri, T. J., Rogers, B. P., Gore, J. C., & Gauthier, I. (2009b). Beyond shape: how you learn about objects affects how they are represented in visual cortex. *PLoS One*, *4*(12), e8405.
Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141-145.
Young, A W, Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-59.
Yovel, G., & Kanwisher, N. (2004). Face perception: domain specific, not process specific. *Neuron*, *44*(5), 889-98.
Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256-62.
Zhang, L., & Cottrell, G. W. (2004). When Holistic Processing is Not Enough: Local Features Save the Day. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.
Zhang, L., & Cottrell, G. W. (2005). Holistic Processing Develops Because it is Good. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.
Zhang, L., & Cottrell, G. W. (2006). Look Ma! No Network!: PCA of Gabor Filters Models the Development of Face Discrimination. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.