Less than Meets the Eye: the Diagnostic Information for Visual Categorization

Jiayu Zhan

Ph.D

School of Psychology

University of Glasgow

May 2019

Abstract

Current theories of visual categorization are cast in terms of information processing mechanisms that use mental representations. However, the actual information contents of these representations are rarely characterized, which in turn hinders knowledge of mechanisms that use them. In this thesis, I identified these contents by extracting the information that supports behavior under given tasks - i.e., the task-specific diagnostic information.

In the first study (*Chapter 2*), I modelled the diagnostic face information for familiar face identification, using a unique generative model of face identity information combined with perceptual judgments and reverse correlation. I then demonstrated the validity of this information using everyday perceptual tasks that generalize face identity and resemblance judgments to new viewpoints, age, and sex with a new group of participants. My results showed that human participants represent only a proportion of the objective identity information available, but what they do represent is both sufficiently detailed and versatile to generalize face identification across diverse tasks successfully.

In the second study (*Chapter 3*), I modelled the diagnostic facial movement for facial expressions of emotion recognition. I used the models that characterize the mental representations of six facial expressions of emotion (Happy, Surprise, Fear, Anger, Disgust, and Sad) in individual observers. I validated them on a new group of participants. With the validated models, I derived main signal variants for each emotion and their probabilities of occurrence within each emotion. Using these variants and their probability, I trained a Bayesian classifier and showed that the Bayesian classifier mimics human observers' categorization performance closely. My results demonstrated that such emotion variants and their probabilities of occurrence comprise observers' mental representations of facial expressions of emotion.

In the third study (*Chapter 4*), I investigated how the brain reduces high dimensional visual input into low dimensional diagnostic representations to support a scene categorization. To do so, I used an information theoretic framework called Contentful Brain and Behavior Imaging (CBBI) to tease apart stimulus information that supports behavior (i.e., diagnostic) from that which does not (i.e.,

nondiagnostic). I then tracked the dynamic representations of both in magneto-encephalographic (MEG) activity. Using CBBI, I demonstrated a rapid (~170 ms) reduction of nondiagnostic information occurs in the occipital cortex and the progression of diagnostic information into right fusiform gyrus where they are constructed to support distinct behaviors. My results highlight how CBBI can be used to investigate the information processing from brain activity by considering interactions between three variables (stimulus information, brain activity, behavior), rather than just two, as is the current norm in neuroimaging studies.

I discussed the task-specific diagnostic information as individuals' dynamic and experienced-based representation about the physical world, which provides us the much-needed information to search and understand the black box of high-dimensional, deep and biological brain networks. I also discussed the practical concerns about using the data-driven approach to uncover diagnostic information.

Acknowledge

I would like to thank many people for their support during my PhD study. First, I would like to thank my supervisor **Prof Philippe G. Schyns** who introduced me to the fascinating field -- visual science, guided me to generate many creative ideas, and opened my mind and trained me to incorporate knowledge and techniques from multiple disciplines; **Dr. Nicola J. van Rijsbergen** who taught me many hands-on skills for data analysis and provided invaluable encouragement and advice throughout my study; **Dr. Oliver G. B. Garrod** who showed lots of patience to answer my questions about programming and provided indispensable technique supports; and **Dr. Robin A. A. Ince** who helped me a lot with his expertise in statistics. I would also like to thank **China Scholarship Council** for providing me the financial support to complete my PhD research; my friends **Dr. Yue Li**, **Dr. Danyang Wang** and **Dr. Yixuan Zhu** for their four years company in Glasgow and the great memories they left.

Lastly, I would like to express my grateful thanks to my parents in China for their unconditional love and endless support in all my pursuits.

Table of Contents

List of Tables

Study 1: Modelling the diagnostic information for Familiar Faces Identification

List of Figures

Study 1: Modelling the diagnostic information for Familiar Faces Identification

List of Supplemental Materials

Study 1: Modelling the diagnostic information for Familiar Faces Identification

Author’s Declaration

I declare that this thesis has been composed by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference, the work presented is my own.

1 General Introduction

Categorization is the must-do task undertaken by our visual system. It is essential to human life because the categorization outcome allows us to make appropriate decisions and survive in the world. For example, we can categorize people as our friend or foe to decide whether we approach or elude, categorize foods as edible or poisonous to decide whether we eat or not, and categorize the environments as safe or dangerous to decide whether we involve or flee.

To deal with the daily categorization task, our brain is proposed as a multi-layered architecture *(Bullmore & Sporns, 2009; K. J. Friston & Kiebel, 2009; Grill-Spector & Weiner, 2014; Guclu & van Gerven, 2015; Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2013; Mumford, 1992; Van Essen, Anderson, & Felleman, 1992)* to transform the high-dimensional information representation mapped onto the retina (e.g. the full face morphology, complexion, and dynamic information) into low dimensional information representations (e.g. selectively process the lip corner puller and the eye wrinkled, see *Schyns, Bonnar, & Gosselin, 2002*) that support subsequent decision (e.g. a happy face). Such a stimuli input to behavioral output transformation along the brain hierarchies implicitly casts visual categorization as an information processing issue. Understanding the mechanisms of visual categorization in cognitive and neuroscience, therefore, requires tracing the information processing that subserves the corresponding behavior.

Before we make endeavours to draw the full map of the information computation in the densely inner-connected, hierarchically organized brain network, we should start our analysis from the information goal: the information that the brain must process to resolve the categorization. Such goal matters as the information requirements from the stimuli input change under different categorizations (*Harel, Kravitz, & Baker, 2014; Schyns et al., 2002; Sigala & Logothetis, 2002*). For example, assigning a face stimulus to a specific identity *Mary* requires different information compared to categorizing the same face as Happy.

My thesis focuses on uncovering such task-specific (diagnostic) information. In this chapter, I will start with ***section 1.1*** to explain in more details why we should start our analysis from diagnostic information. Then in ***section 1.2***, I will discuss the appropriate methodology to uncover the diagnostic information. In

section 1.3, I will review the research topics I'm interested in: ***face identification***, ***facial emotion recognition***, ***neural representations of visual categorization,*** and discuss the unsolved issues from the aspects of diagnostic information representation.

1.1 Diagnostic Information for Visual Categorization

Human visual categorization refers to the cognitive process where the observer assigns the visual input received by the retina to discrete categories, according to their knowledge *(Bar et al., 2006; Goldstone, 1994; Schyns, Goldstone, & Thibaut, 1998; Summerfield & de Lange, 2014)*. Computationally, we can translate the visual categorization process as such: human observers associate the visual stimulus X_i to a category label Y_i via a transformation function $f(X_i) \approx Y_i$. If we implement such input to output transformation in the human brain, it introduces two essential aspects: 1) Stimulus Representation: how does our visual system represent the visual input X_i (i.e., the representational basis)? 2) Transformation Function: how our brain uses the stimuli representations to make correct categorization (i.e., the algorithm defined by the function *f*)? A complete model of visual categorization requires the specifications of both: the transformation function and its representational basis.

In this section, I will firstly review the visual categorization research from the aspects of representational bases and their transformation in the fields of biological vision, computer science (see ***1.1.1***), and cognitive psychology (see ***1.1.2***). Based on these reviews, I discuss the necessity to take the diagnostic information as a critical component (the information goal) to fertilize each field and to guide the complete information processing explanation of visual categorizations (see ***1.1.3***).

1.1.1 Visual Categorization in Biological and Computer Vision

How the human visual system represents the stimuli input along the visual hierarchies is mainly investigated by biological vision. It has been widely demonstrated that visual representation forms a spectrum ranging from low-level to intermediate-level and then to higher-level according to the constraints of representation at different hierarchies of the visual system. For example, at low-level, representational models of primary visual cortex (V1) are mainly based on the output of multi-scale, multi-orientation Gabor filters, which mimics the turning

function of cortical cells in these early regions that is sensitive to different orientations and spatial frequencies *(Hubel & Wiesel, 1998; Jones & Palmer, 1987)*. At intermediate levels, the low-level outputs are summarized and combined to describe more complex features *(Riesenhuber & Poggio, 1999; Rodriguez-Sanchez & Tsotsos, 2012)*, e.g., curvature or local patches (see reviews, *Kubilius, Wagemans, & Op de Beeck, 2014; Peirce, 2015*), in intermediate level brain regions such as V2 (*Freeman & Simoncelli, 2011; Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013*) and V4 (*Pasupathy & Connor, 1999, 2001, 2002; Yamane, Carlson, Bowman, Wang, & Connor, 2008*). The intermediate level outputs are then fed into the higher visual hierarchy (e.g., temporal cortex) to form features that capture higher complexity (*L. Chang & Tsao, 2017; Huth, Nishimoto, Vu, & Gallant, 2012; Kornblith, Cheng, Ohayon, & Tsao, 2013; Kourtzi & Kanwisher, 2001*) and finally enable the categorization.

In computer vision, many hierarchical models (*Fukushima, 1988; Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007; Ullman, 2007*) and deep learning architectures (*Guclu & van Gerven, 2015*, and see reviews *Kriegeskorte, 2015; LeCun, Bengio, & Hinton, 2015*) have algorithmized this low-intermediate-high level representations and their transformation with outstanding categorization performance. As they are beyond my knowledge and not the focus of my thesis, I will not review these models and their algorithms in details here; instead, I would like to raise a conceptual concern which I now detail in next.

There is one challenge associated with the stimuli representations and their operations along the visual hierarchy. On the one hand, the representations across different level must capture sufficient detail so that the outcome of the higher-level contains enough features to allow accurate categorization (i.e., the accuracy concern). On the other hand, as the lower-level representations flow into higher visual hierarchies to create complex representations, robust diagnostic recognition is required to transfer the ‘correct’ features and leave out the noise; otherwise, the process will be highly resource-intensive (i.e., the efficiency concern).

However, the accuracy-efficiency trade-off does not attract enough attention in the field of biological and computer vision. In biological vision, researchers attend mainly to the ‘perceptual’ component, so they emphasize the parameters that are constrained by neurobiological data and aim to design hierarchical models

that are biologically feasible. However, they overlook the ‘cognitive’ component; most of these models are feed-forward so that they learn all the information in the visual input regardless of their diagnosticity for the categorization task at hand. This neglect arguably weakens the validity of these models as they incorporate information that could be redundant to human observers.

In computer vision, the machine learning boom has resulted from neural networks with deep learning architectures and many deep networks recruit the back-propagation algorithm that propagates an error signal from more superficial (categorization or higher-level representation) layers to deeper (perceptual representation) layers. In such recurrent structures, only visual features that are diagnostic for categorization tasks will be learned, but they are diagnostic to the machine learning/algorithmic implementation. As the human brain and machines implement differently by nature, the diagnostic features learned by machine by minimizing the error signal between each layer is demonstrably not the same as the actual diagnostic features learned by human observers in reality (*Nguyen, Yosinski, & Clune, 2015; Phillips, Hill, Swindle, & O'Toole, 2015; Phillips & O'Toole, 2014*).

Understanding the visual representation with biological constraints while considering the accuracy-efficiency trade-off, therefore, requires the incorporation of cognitive constraints. Cognitive constraints are investigated a lot by the psychological research on human observers, which I now introduce in the next section.

1.1.2 Visual Categorization in Cognitive Psychology

Three categorization models[1] are popular in the field of psychology to understand the cognitive aspects of categorization, They are *prototype*, *exemplar*, and *boundary models* (see more categorization models discussed in *Kruschke, 2008*). Prototype models propose that individuals calculate the similarity between the input stimulus and the prototypes (or best example) they have memorized for each category, and respond to the category that has the highest similarity (*Reed, 1972*). Exemplar models assume that individuals calculate a global similarity between the input stimulus and every exemplar of each category, and choose a

[1] I do not aim to discuss every categorization models and their variants in details. I list three main ones here to derive three components related to the representational basis and transformation function.

response based on the global similarity (*Estes, 1986; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986*). The boundary model, which does not explicitly specifies the contents (e.g., the prototype or every instance) for categorization, describes the decision boundaries along information dimensions to dissociate different categories through learning (*Ashby & Gott, 1988; Ashby & Maddox, 1992*). For example, an observer might use the criteria 20-floors tall as a boundary between skyscraper and non-skyscraper, i.e., categories a building as a skyscraper if it is at least 20-floors tall.

Though the categorization processing is implemented differently in these models, we can derive three main components from them: 1) *stimulus representation*, 2) *mental representation* and 3) *decision making*. The *stimulus representation* in cognitive psychology is always specified in a different form than the biological vision. It does not describe stimulus in its biologically-constrained form (see my discussion in ***1.1.1***); instead, it represents stimulus as a point in a multidimensional space in terms of their physical attributes, with the axes of the space representing features, and the prototype at the center of the space. The physical similarity between two items is described as their distance in the space (e.g., *Blanz & Vetter, 1999* for face; *Murase & Nayar, 1995* for object; see recently *C. H. Chang, Nemrodov, Lee, & Nestor, 2017; Nestor, Plaut, & Behrmann, 2016* for the space reconstruction based on human behavioural and neural data). The *mental representation* stored in memory is defined differently in the three models. In the prototype (vs. exemplar) model, the representation of prototype (vs. all know individual exemplars) in the space is memorized; in the boundary model, the feature dimensions in the space that maximize the categorization accuracy are learned and memorized. To transform the input stimulus to a category output, observers compare the *stimulus representation* with their *mental representation*, calculate their similarity (e.g., Euclidean or city-block distance in the multidimensional space) and, if they are close enough, *decide* a category.

Mental representation is the critical component during input-to-output transformation in categorization models: it describes the information human observers do actually have and extract from stimulus representation before their response can be made for a particular task under the model. It explicitly deals with the content that is overlooked by biological/computer vision. However, in these models, the hypothesized representational contents could be too redundant (e.g.,

in the exemplar model), too general (e.g., in the prototype model) and even falsely hypothesized and impair the categorization (see the limitation of *Traditional Hypothesis Approach* in later section ***1.2.2.1***). Precise access to the mental representation requires more careful estimation of the information that determines the categories.

1.1.3 Diagnostic Information Matters

Based on my discussions in ***1.1.1*** and ***1.1.2***, we know that models of visual categorizations in both biological/computer vision and cognitive psychology have their limitations.

Models in biological and computer vision focus on describing how the human brain represents the input stimuli along the visual hierarchy with biological feasibility (representational basis), and what algorithms can optimize (e.g., by minimizing the error) the mapping between input and output (transformation). However, they overlook the impact of humans' top-down knowledge so that the modelled and learned information by machines could be suboptimal and misleading to the human mind.

Models in cognitive psychology release the stimuli representation from the biological constraint and build them in a more psychological traceable space (representational basis). In such space, they modelled the input-to-output transformation, using a critical component *mental representation* to determine what input information should be used based on observers' top-down knowledge. Thus, these models theorize the missing component in biological and computer vision. However, these models are with much disagreement on what actual information the mental representation characterizes, and we cannot derive a complete information processing explanation unless we have precise access to these representational contents.

Now, I would like to take Diagnostic Information as a critical component for visual categorizations, and we need to investigate it explicitly in addition to the Stimulus Representation and the Transformation function that I formulated at the very beginning. From a functional standpoint, diagnostic information is the information that decimates the categories in principle, so it defines the information that necessarily connects a stimulus input to a category output. As such, diagnostic information provides a much precise estimation of the mental

representation. From a practical standpoint, as diagnostic information specifies the information contents of mental representation, it sets up the much-needed information for a task and enables the search of the emergence of behavior from the haystack of neural and (neutrally-inspired) activity. For example, by tracing the diagnostic information processing in the brain, we can track in the brain how the low-to-high feed-forwarding representation along the visual hierarchy is trimmed by a top-down propagation.

In the next section, I will introduce how we can infer diagnostic information from human behavior.

1.2 Inferring the Diagnostic Information from Human Observers

1.2.1 The Nature of Diagnostic Information

To categorize a visual input, the observer extracts task-relevant (i.e., diagnostic) information based on their knowledge stored in memory. There are two facets to this:

1) It represents only a subset of the full information space (*Schyns et al., 2002; Sigala & Logothetis, 2002*), which can be partly predicted from a consideration of the task goal (e.g. a subset of face morphology/complexion for individual identification, or a subset of facial movements for dynamic expression recognition). Such diagnostic recognition increases the coding efficiency and releases the brain from the burdensome computation to achieve the perceptual goal.

2) It reflects observer's prior knowledge (*Jack, Blais, Scheepers, Schyns, & Caldara, 2009; Jack, Garrod, Yu, Caldara, & Schyns, 2012*) and therefore is difficult to predict – idiosyncrasies may reflect organism goals/bias unrelated to the explicit task.

All these aspects of diagnostic information influence the experiment design to reconstruct the diagnostic information from the human mind.

1.2.2 Deriving the Diagnostic Information

At least two prerequisites we need to take into consideration:

1) diagnostic information comprises only a subset of full features, but uncertainty about the subset requires the testing to cover the full information space as much as possible.

2) as diagnostic information can be idiosyncratic and reflects observers' own prior experience, it requires experimenters to set their own knowledge aside as far as possible to allow an unbiased estimation.

However, the typical hypothesis-testing approach used in psychology is limited to satisfy the prerequisites (see ***1.2.2.1***), and we should adopt a Data-driven approach widely used in psychophysics for a broader and unbiased investigation (see ***1.2.2.2***).

1.2.2.1 Traditional Hypothesis Testing Approach and Limitations

Typically, experimenters set their hypothesis as "feature X elicits the processing of category Y." To test the hypothesis, experimenters always show a series of stimuli to participants, with feature X present or absent, and ask them to categorize each according to a set of categorical labels. Based on participants' response, e.g., the presence of feature X is always related to the response label Y, experiments attribute the factors that drive the categorical decision Y to the feature X. However, I argue that hypotheses formulated in such a way are not sufficiently powerful to uncover the diagnostic information, at the level of details and precision.

The first shortcoming of the typical design is the lack of (or poorly defined) content of the hypothesis. For example, one experiment showed that participants tend to perceive the anger faces of males or Caucasians as more dominant than faces of female or another ethnicity (*Hess, Blairy, & Kleck, 2000*). This result simply associates one stimulus category (i.e., male or Caucasian) with higher perceived dominance ratings than another stimulus category, thus it cannot specify what information present in the face (or particularly in the male or Caucasian faces) drives the perception along the dimension of dominance. Another example is the face inversion effect. Researchers found that participants' performance on face identity recognition is severely impaired when faces are displayed in their inverted orientations (*R. K. Yin, 1969*), and they ascribe such effect to a holistic processing of face information, i.e., a processing that integrates the facial features into a gestalt whole (*Taubert, Apthorp, Aagten-Murphy, & Alais,*

2011). However, the holistic processing is poorly defined: what actual facial features are integrated to form a gestalt whole?

The second shortcoming is that the typical design is less sensitive to participants' idiosyncrasies. An example is a well-known hypothesis about the universal recognition of six facial expressions of emotion, i.e., happy, surprise, fear, disgust, anger, and sad, which however turned out to be inappropriate. The universality hypothesis is inspired by Darwin's theory which suggested these basic facial expressions have an evolutionary and biological basis. By virtue of such origins, researchers considered these signals should be similar among humans so that can be recognized regardless of culture (*Ekman et al., 1987; Izard, 1994*). To test this hypothesis, Western researchers derived a small set of facial movement patterns to represent six emotions according to their theory (*Ekman, 1971*) and asked both Western and non-Western participants to categorize these stimuli using a six-emotions alternative-force-choice task (6 AFC, see pioneer research *(Ekman, Sorenson, & Friesen, 1969)*). Indeed, both groups of participants can categorize (above chance level, 16.67%) these theoretically-derived models to their corresponding hypothesized emotion labels in the 6 AFC task. However, if we look at the data carefully, the non-Western societies showed much lower recognition level (*Elfenbein & Ambady, 2002; Jack et al., 2009; Nelson & Russell, 2013*), suggesting cultural specificities of facial emotional signals. A recent data-driven investigation demonstrated such culture-specific knowledge about the facial expressions of emotion (*Jack, Garrod, et al., 2012*).

In the next section, I will introduce how we can use the data-driven approach to overcome the above limitations.

1.2.2.2 Reverse Correlation Approach

Based on the discussions in ***1.2.2.1***, an appropriate hypothesis should be formulated to test on rich information spaces. The reverse correlation approach used in psychophysics is a good practice (*Murray, 2011*).

Reverse correlation approach breaks down stimuli information by sampling them parametrically in an information space, which increases the level of granularity and range in the content we can test. For example, to uncover what information present in the face that drives the perception of dominance, we can model the face as their position in a multidimensional space. The axes of the high-

dimensional space represent the physical attributes of the face with high resolution. To sample face features covering a broader range, we can create many face stimuli by parametrically setting their weights on the multi-dimensions of the space. The testing is then performed on the fine-grained sampled contents.

Reverse correlation approach measures the perceptual decision on the randomly sampled information, which releases our testing from a predefined subset of information (e.g., the X features I mentioned above) and increases the power to capture participants' idiosyncrasies. Let me use the facial expression signals to illustrate. Rather than theorizing a small set of facial movement patterns (e.g., the models proposed under the universality hypothesis), researchers can generate a face with a subset of randomly selected facial movements displayed. Many such random samples can constitute a set of stimuli that thoroughly cover the full information space of natural facial movements, which are then tested against a spectrum of perception (6 emotions plus 'don't know'). The rationale is that observers can only categorize the sampled facial movement signal as 'happy' when it comprises diagnostic signals of the 'happy' based on their mental representation (e.g., eyes wrinkled, cheek raised, and lip corner puller); in contrast, when the sampled information does not contain diagnostic information for any of the categories, the observer will respond 'don't know'.

By measuring the spectrum of the perception that rich stimuli sampling produces, reverse correlation approach can create a transfer function to describe how different information samples of the stimuli contribute to different categorizations, and based on which experimenters can infer the task-specific diagnostic information. For example, experimenters can measure the relationship between the sampled stimuli (i.e., the visibility of different facial movement) and the corresponding perceptual spectrum measuring (i.e., six emotions plus 'don't know'), using correlation, linear regression, or information theory. The resulted statistical parameters (e.g., *r*-value in correlation, β coefficients in regression, and mutual information in information theory) can quantify the contribution of each facial movement to the perception of different emotions.

Critically, the reverse correlation approach is generic. We can adapt it to any types of visual stimuli that sampled in a variety of ways (see a review paper *Jack & Schyns, 2017*), to any categorization task, and test on any participant groups. Such a broader and more rigorous exploration uncovers the critical

information that drives perceptual decisions which might be hidden by experimenter's own knowledge.

In my thesis, I applied reverse correlation approach to uncover the diagnostic information for face identity recognition and facial expression recognition, and to investigate further how human brain reduces the stimuli representations to the diagnostic representation for perceptual decisions. Given the focus of my thesis, I will review the research in each field in next sections to obtain a clear picture of the theoretical and empirical status and to understand how the investigations from the diagnostic recognition can contribute to these fields.

1.3 Topics of Interest

1.3.1 Face Identification

Humans can remember hundreds of individual faces and identify each amongst others effortlessly under various conditions of pose, illumination, and ageing. This suggests the encoding and representations of face information that makes the face identification unique. There are mainly two lines of research in field of face identification.

1.3.1.1 Holistic Processing of Faces

Researchers interpreted the speciality of human face processing, compared with non-face categories, as its holistic way of representation; that is, face is not represented in its isolated component parts (e.g. round eyes, thin nose, and pouty mouth) or their combinations but as an integrative and un-decomposed whole. Empirical evidence supporting the holistic processing is from the behavioural tasks using inverted faces (*Carey & Diamond, 1977; Freire, Lee, & Symons, 2000; Valentine & Bruce, 1986; R. K. Yin, 1969*), composite faces *(Rossion & Boremanse, 2008; Young, Hellawell, & Hay, 1987*), and the part-whole recognition (*Tanaka & Farah, 1993*). In the face inversion effect, participants show more difficulties to recognize faces represented upside down than in their upright positions. As such inversion disrupts less for the objects *(Rossion et al., 2000*) and the isolated face parts recognition (*Rhodes, Brake, & Atkinson, 1993*), researchers attributed the impaired performance of inverted face (vs. objects or face parts) to the nature of global-based (vs. part-based) processing. In the composite face task, researchers assembled upper and lower half-faces from two identities, and they

found participants are slower to recognize the identity of a half-face when the two halves are aligned than they are misaligned. In the part-whole task, participants performed better to recognize a face part (e.g. nose) of a target identity when it is presented in the original face than in isolation. Both the composite and part-whole effects indicate that the representation of individual face feature is interfered by the presence of other parts of the face, suggesting a mandatory processing of features integration as a whole.

1.3.1.2 Feature-based Representations in Face Space

In the feature-based approach, researchers investigated the face representations using the 'face space' model. To create the face space, researchers used 2D face images or 3D faces and applied dimensionality reduction techniques (e.g. PCA see *Turk & Pentland, 1991* and multidimensional scaling) to formalize a multidimensional space, where each dimension defines a face feature (e.g. an eigenface or classification image). Thus, each face has a weight on each feature dimension and its position in the space reflects how it can be represented as the combination of these component features. The face space framework offers an efficient coding scheme by referencing the quantities and qualities of variability in the population of human faces, rather than specifying the 'unique' identity information in an absolute term.

The face space opens the door to investigate whether and how the features derived from objective face information can account for human subjective face identification. For example, to understand the contribution of each feature dimension to memory representations (including their neural coding), researchers modelled the relationship between the project weights of original 2D face images on each dimension and participants' corresponding behavioural *(C. H. Chang et al., 2017*) and brain (*H. Lee & Kuhl, 2016; Nestor et al., 2016*) response. As shown by Chang and Tsao (*2017*), neurons selectively respond along a single axis of the face space, not to the other, orthogonal axes, suggesting the feature-based identity representations in the brain (i.e. the axis model). In another line of study, researchers altered the feature weights of a target face in the face space into its opposite and create its anti-face. They found the adaption to the anti-face can bias the participants' perception towards the space centre, i.e. they recognized the average of the target face and centre face in the space as the original target. Such centre-shift aftereffect supports the norm-based coding theory of face identity:

human observers represent each face identity according to how they derive from the average of a multi-dimensional face space *(Leopold, O'Toole, Vetter, & Blanz, 2001; Rhodes & Jeffery, 2006*). Monkey single cell responses show increased firing rate with increasing distance of a face to this average (as happens with e.g. caricaturing the feature values, *Leopold, Bondar, & Giese, 2006*), supporting the norm-based representation at the neural level.

1.3.1.3 Unsolved Issues

Under the diagnostic recognition framework, diagnostic information constitutes only a subset of the full information space based on the task goal (c.f. ***1.2.1 The Nature of Diagnostic Information***). In the holistic processing approach, the researchers manipulated the face features in an arbitrary way, which cannot specify the source of the holistic or integration effect, i.e. what actual features are represented for integration and how they are integrated? Understanding the holistic processing of human face therefore requires the content-based emphasis. In the face space model, though it quantifies the face information and enables the well-controlled manipulation, the analysis typically adopts a brute force approach: it explains the variance comes from physical face shape and texture information from an average which can over fit subjective representation of (a subset of) face shape and texture. Thus, there is no provision in such physical face faces to enable better recognition of some faces (familiar faces) than others (unfamiliar faces). In sum, the investigation builds on the use of diagnostic information is always neglected in facial identity processing.

1.3.2 Facial Emotion Recognition

Mutual understanding of emotion between individuals is critical to human interactions, achieved primarily by exchanging a set of facial expressions. Accurately recognizing facial emotions requires the shared representation of the expressions in the mind between signal sender and receiver (*Jack & Schyns, 2015*), which allows individuals to use the same set of diagnostic information to understand each other's state, demands and intentions, and coordinates them to behave optimally in the physical and social environment. Given the fundamental function of facial emotion in human society, understanding the mental representation of face emotions across individuals has been a primary goal in visual science.

1.3.2.1 Representation of Facial Expressions are Shared across Individuals

Based on Darwin's theory, the true origin of facial expressions comes from their adaptive functions to increase the chance to survive (*Darwin, 1999*). For example, if we look at a disgusted face, it is typically characterized by the wrinkled nose and squeezed eyes. Such kind of facial muscle contraction can protect us against the exposure to noxious contaminants and signal companions about the threatening environment. By natural selection, the facial signals are then passed onto the next generation for survival. Under such an evolutionary view, facial expression of emotions should be innate, i.e. they are hardwired before the birth rather than learned during social interactions. Thus, representations of facial expressions should be shared across individuals.

To provide empirical evidence to support the 'universe hypothesis', Ekman and his colleagues did a series of pioneer studies, using a standard set of theoretically-derived models (i.e. prototypes) that they proposed to capture observers' representation of six facial expressions of emotions (*Ekman & Friesen, 1978*). In a classic study, Ekman and his colleagues selected a set of facial photographs that display the prototypical facial expression of six emotions, then they tested the recognition of these photographs using a 6 emotions alternative-force-choice (AFC) task in the observers from 5 different cultures (i.e. United States, Brazil, Japan, New Guinea, and Borneo). They found high agreement (i.e. based on above chance accuracy) across all cultures, and concluded a universal representation of facial expressions of six emotions. Since then, researchers start to use the standard model set for six emotions developed by Ekman to test human facial emotion recognition and to create facial expression databases (e.g., Rodbound faces database*, Langner et al., 2010*; and Japanese female facial expressions database*, Lyons, 1998*).

1.3.2.2 Representation of Facial Expressions are Shaped by Culture

As more and more data are collected in various cultures, researchers start to challenge the validity of the standard facial expressions developed by Ekman as the universal models characterizing the mental representations of six emotions in observers across all culture (see a review *Russell, 1994*; and a meta-analysis study *Elfenbein & Ambady, 2002*). One criticism is to use nAFC task (e.g. n = 6 where the 6 emotion labels are used) to test the universal hypothesis. In a typical

nAFC design, experimenters selected a limited set of response labels in a top-down manner which can misrepresent participants' perception (c.f. my discussion in ***1.2.2.1***), and the force-choice therefore lead the response to an ill-fitting category, especially in the absence of an 'other' option (see also the discussion in *Russell, 1993*). Also, researchers typically used the proportion of correct response (based on the "universal model") above chance level to determine whether a testing model is recognizable or not. However, without the estimation of response bias (uniform vs. modal distribution of response across pre-selected categories), experimenters simply comparing the proportion of correct response with chance-level can increase Type I error and weaken the validity of conclusion they can draw *(Jack, 2013)*.

In a facial expression production experiment *(Elfenbein, Beaupre, Levesque, & Hess, 2007*), researchers directly asked participants to pose emotional expressions that they thought "their friends would be able to understand easily what they feel." The results showed that participants from two different culture groups (Quebec vs. Gabon) used different facial movements to produce "happiness", "surprise," "anger" and "sadness". The authors further tested these culture-specific posed emotional expressions to a new group of participants and found an in-group advantage, i.e. participants recognize the emotional expressions from their own culture significantly better than the prototypical models proposed by Ekman and Friesen (*Ekman & Friesen, 1978*).

Using a data-driven approach, Jack and her colleagues directly modelled the mental representations of 6 facial emotions, separately for Western and East Asian participants *(Jack, Caldara, & Schyns, 2012; Jack, Garrod, et al., 2012*). In one study, they added random white noise (i.e. grey-scale pixel values) to a neutral face image to randomly change the appearance of the face. Then they asked participants to make a 7AFC task (6 emotions plus 'don't know') on these stimuli. If the noise changes the neutral face in a way that matches the mental representation of facial expressions of the observers, they will categorize the neutral face as the corresponding expressive signal. To visualize the representational contents of each emotion, researchers averaged the noise in the trials associated with each emotion and then added it to the neural base face. The resulting contents (i.e. the averaged noise template) showed that, to perceive the neutral face as expressive, Westerners required the information to be added to the

eye and mouth regions whereas Asian people required more information to be added to the eye regions to perceive emotions (*Jack, Caldara, & Schyns, 2012*).

As the facial expression is by nature dynamic, in a later study they created a series of random facial movements and tested them against 7 categorical options (i.e. 6 emotions plus 'don't know'). By analysing the relationships between the random facial movements and corresponding categorization behavior, they derived a set of dynamic models representing the mental representations of each emotion, independently for Western Culture and Eastern Cultural groups. By analyzing the facial movement patterns and their temporal dynamics, they demonstrated culture-specific representations for facial expressions of emotions: 1) Westerners use distinct movement pattern to represent each emotion but Easters do not; 2) Easterners rely on early eye activity to represent emotion intensity whereas Westerners represent emotional intensity using other face parts (*Jack, Garrod, et al., 2012*).

1.3.2.3 Unsolved Issues

Under the diagnostic recognition framework, diagnostic information reflects observers' memory representation based on their past experience (c.f. ***1.2.1 The Nature of Diagnostic Information***). Different cultures are formed according to how the culture members interact with the physical world and in return shape their conceptual knowledge (i.e. mental representation) about the visual environment. Therefore, the standard set of mental models derived from the hypothesis-driven approach in one culture is ill-fitted the mental representations of individuals in other cultures. The valid set of expression models should be created for each culture, independently.

Less bounded by experimenters' prior knowledge, the mental models reconstructed in the reverse correlation approach should depict participants' mental representation more precisely in principle. However, if we look at the procedure closely, each model is created by measuring across trials the relationship between Random Facial Movements and 7 Categorization, using the response of one specific participant (see methodology detailed in *Jack, Garrod, & Schyns, 2014 and Jack, Garrod, et al., 2012*). Thus, each model characterizes the mental representation of the tested participant, not the whole culture group, and need to be validated for its generalizability – i.e., the model is recognizable to other observers when displayed on different faces.

1.3.3 Visual Categorization in General

For face identification, we see identity information from a face; for facial expression recognition, we see emotional information from a facial movement pattern. Now I would like to extend my reviews to the general visual categorization, i.e. to determine what we see from the visual input. In the literature, visual categorization, recognition, and visual perceptual decisions are always used interchangeably, which cover a broad range of stimuli from very basic types (e.g. dots motion, line orientation) to more complex ones (e.g. face, objects and scene). In my thesis, when I discussed visual categorization, recognition, or perceptual decisions, I refer to the same process by which our brain gathers and integrates information from the visual stimuli input and assigns it to a categorical proposition. I focus mainly on the research about complex stimuli, including face, object and complex scene recognition.

1.3.3.1 Three Main Processing Components

In a quantitative approach, perceptual decision is thought as a form of statistical inference *(Kersten, Mamassian, & Yuille, 2004; Rao, 1999; Tenebaum & Griffiths, 2001*). Researchers decomposed the inference procedure into three main processing stages: 1) representations of the visual information (i.e. the evidence), 2) accumulating and integrating the evidence across time according to prior knowledge, 3) comparing the calculation output of 2) to a decision threshold to make response (see reviews in *J. I. Gold & Shadlen, 2007*).

The evidence in general refers to the information we rely on to make the inference. For visual categorization, information conveyed by the stimulus is the evidence (e.g. a facial muscle movement such as lip corner puller) we use to make decisions. To use the evidence *x*, we need to interpret it under a given option *R* (for example, expressing a happy face), with the idea of 'likelihood' $p(x \mid R)$. When the relative likelihood of x given the option R (e.g. lip comer puller in a happy face) is over the other ¬R (e.g. lip corner puller in other expressive faces) and above a threshold, we make a decision (e.g. recognize the happy face). Visual categorization always relies on many features but not single one, therefore in most cases the likelihood ratios of multiple evidence (e.g. lip corner puller and the eye wrinkled) are calculated and accumulated over time, and the decision is made when the accumulated ratios supporting one option is above threshold (*Link & Heath, 1975; Ratcliff, 1978; Usher & McClelland, 2001*).

1.3.3.2 Neural Signatures

I review the neural representations of visual categorization by incorporating the 3 decision components discussed in ***1.3.3.1*** into two lines of research. In the first line, researchers focus more on the perceptual component that deals with 'what' and 'where' questions, i.e. coding selectivity of different visual information in different brain regions. This line of research informs the neural representation of visual evidence for different categories. In the second line, researchers investigate the 'how' question under the theoretic framework in ***1.3.3.1***, i.e. relating the brain activity of different regions to evidence accumulation and decision making.

1.3.3.2.1 ***Representation of visual evidence for different categories***

Visual categorization is neutrally implemented in a series of cortical regions, starting from the primary visual cortex (V1) that receives the information mapped on the retina. The primary visual cortex then projects, along the ventral stream, to V2, V3 and V4 - where the low-level physical properties such as contrast or orientation are processed to form texture and contour information *(Freeman & Simoncelli, 2011; Freeman et al., 2013; Pasupathy & Connor, 1999, 2001; Yamane et al., 2008*). These adjacent visual areas have a reversed representation of the external visual world due to the retinotopic mapping, i.e. flipped representation around the vertical (left vs. right) and horizontal (up vs. down) meridian of the visual areas. The visual information then progresses further to the high-level region -- ventral temporal cortex – where represents both the contralateral and ipsilateral of the visual fields and where the category selectivity emerges (see reviews *Grill-Spector & Weiner, 2014; Op de Beeck, Haushofer, & Kanwisher, 2008*).

Face. Many studies using fMRI have discovered that the regions in ventral temporal cortex are activated stronger by face than non-face stimuli. In an fMRI study, Kanwisher and her colleagues (*Kanwisher, McDermott, & Chun, 1997*) localized one region in right fusiform gyrus that was consistently activated for faces than objects across participants in a passive view task. This region also showed face specificity in the tests that used the front-view face vs. scrambled face, front-view face vs. front-view house, and three-quarter-view face vs. human hands contrasts. In addition to the fusiform face area (FFA), a region in the inferior occipital gyrus - the occipital face area (i.e. OFA, *Gauthier et al., 2000; Pitcher, Walsh, & Duchaine, 2011; Pitcher, Walsh, Yovel, & Duchaine, 2007*), and a region

in the posterior superior temporal sulcus (STS, *Haxby, Hoffman, & Gobbini, 2000*) are also found for face-selective representations. The OFA is thought for face detection (*Fairhall & Ishai, 2007; Haxby et al., 2000*), the FFA plays a central role in individual identification *(Gauthier et al., 2000; George et al., 1999; Grill-Spector, Knouf, & Kanwisher, 2004; Winston, Henson, Fine-Goulden, & Dolan, 2004*), and the STS is involved in the recognition of facial emotion and gaze and has been suggested to process dynamic facial information *(Andrews & Ewbank, 2004; Calder & Young, 2005; Hoffman & Haxby, 2000*).

Object. By comparing the brain activity when participants passively viewing the photographs of everyday objects and visual textures, Malach et al. *(1995*) found an object-selective activation located laterally to the fusiform gyrus. In a following fMRI study, Kanwisher et al. *(1996*) found a region, which locates very close to where Malach reported, responds significantly stronger to 3D objects line drawings than to scrambled line drawings. As more neural data collected using the contrast of intact objects vs. control stimuli without clear shapes, researchers proposed an object-selective areas which locate mainly in lateral occipital cortex and extend anteriorly and ventrally into posterior temporal regions (see review *Grill-Spector, Kourtzi, & Kanwisher, 2001*).

Scene. Another category selectivity region is in parahippocampal cortex, which responds stronger when subjects view topographical scene stimuli (e.g. outdoor and indoor scenes) than they view various nonplace controls (e.g. scrambled scenes or faces, *Aguirre, Zarahn, & D'Esposito, 1998; Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998; Nasr et al., 2011*). Due to its specificity for 'place', researchers named this region as parahippocampal place area (PPA). To understand the scene-selectivity of PPA in a causal way, Mégevand et al. (*2014*) stimulated the activity of this region via the intracranial electrodes and found the stimulation can induce a topographic visual hallucination, i.e. the patient reported to see indoor or outdoor scene when he did not perform any task.

Modular vs. Distributed. Until now, I introduced the results showing the category selectivity in discrete regions, i.e. the modular representation. However, the category information is organized in a hierarchical structure: basic-level (e.g., faces vs. cars), superordinate-level (e.g., animate vs. inanimate object), and subordinate-level (e.g., Mary vs. Beetle). Such organization requires the flexible

access to different levels of categorical information. Then the question for module view arises: how many category-specific modules should exist? This is not the topic of my thesis, but I would like to provide some evidence supporting the distributed representation of category-specific information in ventral-temporal cortex. Using the multivoxel pattern analysis, researchers found there is not just one distinct region selective for face, but rather a series of sparsely-distributed clusters along the occipito-temporal sulcus and in fusiform gyrus *(Weiner & Grill-Spector, 2010*). The FFA, face-selective module, has also been demonstrated to process not only face information but also body parts(Weiner & Grill-Spector, 2010), vehicles and animals *(Cukur, Huth, Nishimoto, & Gallant, 2013; Grill-Spector, Sayres, & Ress, 2006; Hanson & Schmidt, 2011; McGugin, Gatenby, Gore, & Gauthier, 2012*).

Either modular or distributed, the ventral temporal cortex which is sensitive to categorical information represents the visual evidence for different categories.

1.3.3.2.2 ***Integration of visual evidence and decision formation***

To directly track the accumulation of the visual evidence, Ploran et al. (*2007*) gradually reveal the contents of the picture on each trail, using a dissolved black mask at each successive 2s interval over the time course. They asked the participants to signal their recognition at any time during the display window by a button press, and recorded their brain activity using fMRI. They found the activities in inferior temporal, frontal and parietal regions are gradually increased as more contents are revealed, and peak at the time corresponding to when the recognition is made. This result suggests these regions accumulate the visual evidence to support the object identity in the picture. In medial frontal cortex, however, the activity remained near baseline until the recognition time, suggesting an operation related to the moment that decision formed.

In the single-unit recording studies in monkeys, researchers found the decision is formed by comparing the response output of lower-level neurons that are sensitive to one category with those sensitive to another category, and such comparison is computed in higher-level cortical regions (e.g. the prefrontal cortex, *Kim & Shadlen, 1999*; and premotor cortex, *Hernandez, Zainos, & Romo, 2002*). To investigate such comparison operation in human brain, Heekeren et al. (*2004*) tested participants' brain response in a face-house discrimination task in the fMRI scanner while they also modulated the task difficulty using the degraded images

(i.e. hard trial) and clear image (i.e. easy trial). Based on the neurophysiological data obtained in monkeys, Heekeren proposed that the comparison computation area should fulfil two criteria: “First, they should show the greatest activity on trials in which the evidence for a given perceptual category is greatest, for example, a greater fMRI response during decision about suprathreshold [clear] images of faces and houses than during decisions about perithreshold [degraded] images of these stimuli. Second, their activity should be correlated with the difference between the output signals of two brain regions containing pools of selectively tuned lower-level sensory neuron involved; that is, those in face- [i.e. FFA] and house-responsive [i.e. PPA] regions.” The only region fulfilled both criteria is the posterior portion of the left dorsal lateral prefrontal cortex, suggesting its role in integrating the response outputs from regions for visual evidence representation and accumulation (e.g. FFA vs. PPA) and using a comparison operation to make decision (face vs. house).

1.3.3.3 Re-evaluation under the Diagnostic Recognition Framework

Relating the neural activities of different regions to the three cognitive components in perceptual decisions enables us to infer the functional architectures of human brain underlying the visual categorization. However, if we look at the literatures carefully, there is still a missing component that reduces the full redundant visual information retinotopically represented in early visual cortex to the category-specific visual evidence represented in ventral temporal cortex. Such information reduction is necessary: to optimize the inference output (i.e. maximizing the percentage of correct response and/or shortening the decision time in the visual categorization task), the represented and accumulated visual evidence should not include redundant (nondiagnostic) information that in principle induces more processing cost, either by decreasing the chance of incorrect response or by increasing the time to accumulate the evidence.

To understand where, when and how the human brain reduces the full stimuli representations to the visual evidence for perceptual decisions, we need to first tease apart visual information that supports category response (i.e. diagnostic information) from that which mapped on the retina but does not support decision (i.e. the nodiagnostic information), and then examine their neural representations separately. Such investigations rely on the reverse correlation approach together

with statistic measurements that can capture the triple relationship between visual stimuli, brain activity and perceptual decisions together.

1.4 Thesis foci

As discussed, diagnostic information plays an important role in understanding the information mechanism of visual categorization; however it does not attract enough attention in literature. With the development of sampling techniques, we can now model the diagnostic information in a much elegant way. In my thesis, I applied novel sampling tools to uncover the diagnostic information for two face categorization tasks: 1) *familiar face identification* (Chapter 2), which taps into the information space of face morphology (i.e. face shape) and complexion (i.e. face texture); 2) *emotional facial expression recognition* (Chapter 3), which taps into the independent dynamic dimension of face information. In Chapter 4, I used a scene categorization task as a case study, to trace where, when and how human brain reduces the stimuli representations to the diagnostic information for perceptual decisions.

2 Study 1: Modelling the Diagnostic Information for Familiar Faces Identification

2.1 Introduction

Observers use their mental representations to identify familiar faces under various conditions of pose, illumination, and ageing, or to draw resemblance between family members. As we always recognize face effortlessly under a variety of conditions, the representational contents must be sufficiently detailed to enable accurate recognition —i.e. identifying 'Mary' amongst other people, and sufficiently versatile to enable recognition across diverse everyday tasks —e.g. identifying Mary in different poses, at different ages or even identifying her brother based on family resemblance (*O'Toole, 2011; Rosch & Mervis, 1975; Tsao & Livingstone, 2008*). And yet, it remains a fundamental challenge to reverse engineer the participant's memory to model and thereby understand the detailed contents of their representations of familiar faces. This challenge is a cornerstone to understand the information processing mechanism of face identification, because they process the contents to predict the appearance of the familiar face of 'Mary' in the visual array and to selectively extract its identity information to generalize behavior across common tasks.

In this Chapter, I studied how our own work colleagues recognize the face of other colleagues from memory. The work environment provided a naturally occurring and common medium of social interactions for all participants, who developed their personal familiarity with the people whose faces the study tested. In ***Experiment 1*** I used a novel 3D face information generator combined with the reverse correlation method to model the 3D face identity information of 4 familiar faces stored in the memory of 14 individual participants. In ***Experiment 2***, I validated these memorized contents in a new group of participants, showing that the these contents compromise the face features that maximally distinguish each identity from a model norm. In ***Experiment 3***, I further demonstrated that the faithful memorized contents (I call them the diagnostic contents) contain information that enables the third group of participants to generalize identification to new tasks with the changes in viewpoint, age, and sex of the face.

2.2 Experiment 1: Modelling the Contents of Mental Representations of Familiar Faces

2.2.1 Participants

I recruited 14 participants (all white Caucasians, 7 females and 7 males, mean age = 25.86 years, SD = 2.26 years) who were personally familiar with each familiar identity as work colleagues for at least 6 months. I assessed participants' familiarity on a 9-point Likert scale, from not at all familiar '1' to highly familiar '9' (see ***2.6.2*** *Supplemental Tables*, *Table S2-1* for their familiarity ratings on each identity). All participants had normal or corrected-to-normal vision, without a self-reported history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (e.g., depression, autism spectrum disorder or prosopagnosia). They gave written informed consent and received £6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

2.2.2 Generative Model of 3D Face Identity (GMF)

My colleagues designed a generative model to objectively characterize and control face identity variance, using a database of 355 3D faces (acquired with a 4D face capture system, see ***2.6.1*** *Supplemental Methods, 3D Face Database*). For each 3D face, its shape is parameterized with the 3D coordinates for each one of 4735 vertices, and its texture is parameterized with the RGB values of 800*600 pixels (see Figure 2-1A). It is critical to reiterate that the familiar faces were not part of the 3D face database.

To design the 3D GMF, we first applied a high-dimensional General Linear Model (GLM), separately to each 3D vertex coordinate and 2D pixel RGB value, to model and explain away variations in face shape and texture that arise from the non-identity categorical factors of sex, age, ethnicity, and their interactions. The GLM therefore: 1) extracted as a non-identity face average the shape and texture face information explained by non-identity categorical factors; and also 2) isolated the residual information that defines the 3D shape and 2D texture identity information of each face--i.e., the identity residuals.

To further control identity information, we applied the Principal Components Analysis (PCA) to the identity residuals of the 355 faces, separately for shape and texture. The PCA represented shape residuals as a 355-dimensional vector in a

355-dimensional space of multivariate components, and a separate PCA represented the texture residuals as a 355 * 5 (i.e. 5 spatial frequency bands)-dimensional matrix in a space of 355*5 multivariate components. Two sets of PCA coordinates therefore represented the objective shape and texture information of each identity in the principal components space of identity residuals.

Our 3D GMF is formally expressed as follows:

$$Faces = Design\ Matrix \times Coefficient\ Matrix + weights \times PCs$$

Where $Faces$ is the vertex (or texture) matrix of 355 faces: for vertices, it is [355 x 14,205] where 14,205 = 4,735 vertices x 3 coordinates; for texture, it is [355 x 1,440,000] where 1,440,000 = 800 x 600 pixels x 3 RBG. $Design\ Matrix$ defined the non-identity categorical factors and their interactions (N = 9), i.e. constant, age, gender, white Caucasian (WC), eastern Asian (EA), black African (BA), gender x WC, gender x EA, gender x BA, for each of face (N = 355), and therefore is [355 x 9]. We estimated the linear effects of each non-identity factor and their interactions using the GLM which are represented in the $Coefficient\ Matrix$ (i.e. [9 x 14,205] for shape and [9 x 1,440,000] for texture). After the GLM fit, the [355 x 14,205] shape (or [355 x 140,000] texture) residuals are further explained using the PCA analysis, resulting 355 components.

Figure 2-1B schematizes the computation flow of the 3D face identity modelling (indicated by solid arrow). The GLM decomposes a scanned 3D face ‘Tom’ into his average face, which captures the non-identity categorical factors of his sex, ethnicity, age and their interactions. The corresponding heat map indicates the left identity residuals of his 3D shape (2D texture, not illustrated in the figure, is independently and similarly decomposed), on which the red color denotes the outward changes of the 3D vertices in relation to his category average whereas the blue color denotes vertices with the 3D inward changes. As shown, Tom has higher cheek bones, wider nose bridge, and flatter brow ridge than his average. His 3D shape residuals are further projected into a multidimensional PCA space to parameterize his specific identity information.

The design of 3D GMF also enables us to synthesize new faces. Figure 2-1B schematizes the reversed computation flow for 3D face identity generation (indicated by the dashed line), using controlled non-identity factors. First, I fitted Jane’s face in the GLM to isolate its non-identity averages; then I created random

identity residuals using random PCA weighs; finally I plus these two together to obtain random face identities shared all other categorical face information with ‘Jane’. I used these generative properties to derive the stimuli used in this experiment (see ***2.2.3.2*** *Stimuli, Random Face Identities*).

Figure 2-2 illustrates the sythesization of new faces, with the controlling of the identity residuals. First, I scanned the four familiar faces of the experiment (2nd column); then, I fitted each into the 3D GMF and got their non-identity GLM averages and identity PC weights. I plus their non-identity averages and their identity weights to generate their ground truth faces (the 3rd column), which show only minimal distortions from the scanned faces(shown in the 1st column). I can also change their non-identity GLM averages, e.g., change their age, sex or ethnicity separately, or jointly sex and ethnicity in GLM, plus them to their identitiy PC weights. The outcomes are older, sex swapped, ethnicity swapped and sex and ethnicity swapped versions of the same identity (the 4th to 7th column). I used these generative properties to derive the stimuli of Experiment 3 (see ***2.4.2*** *Stimuli*).

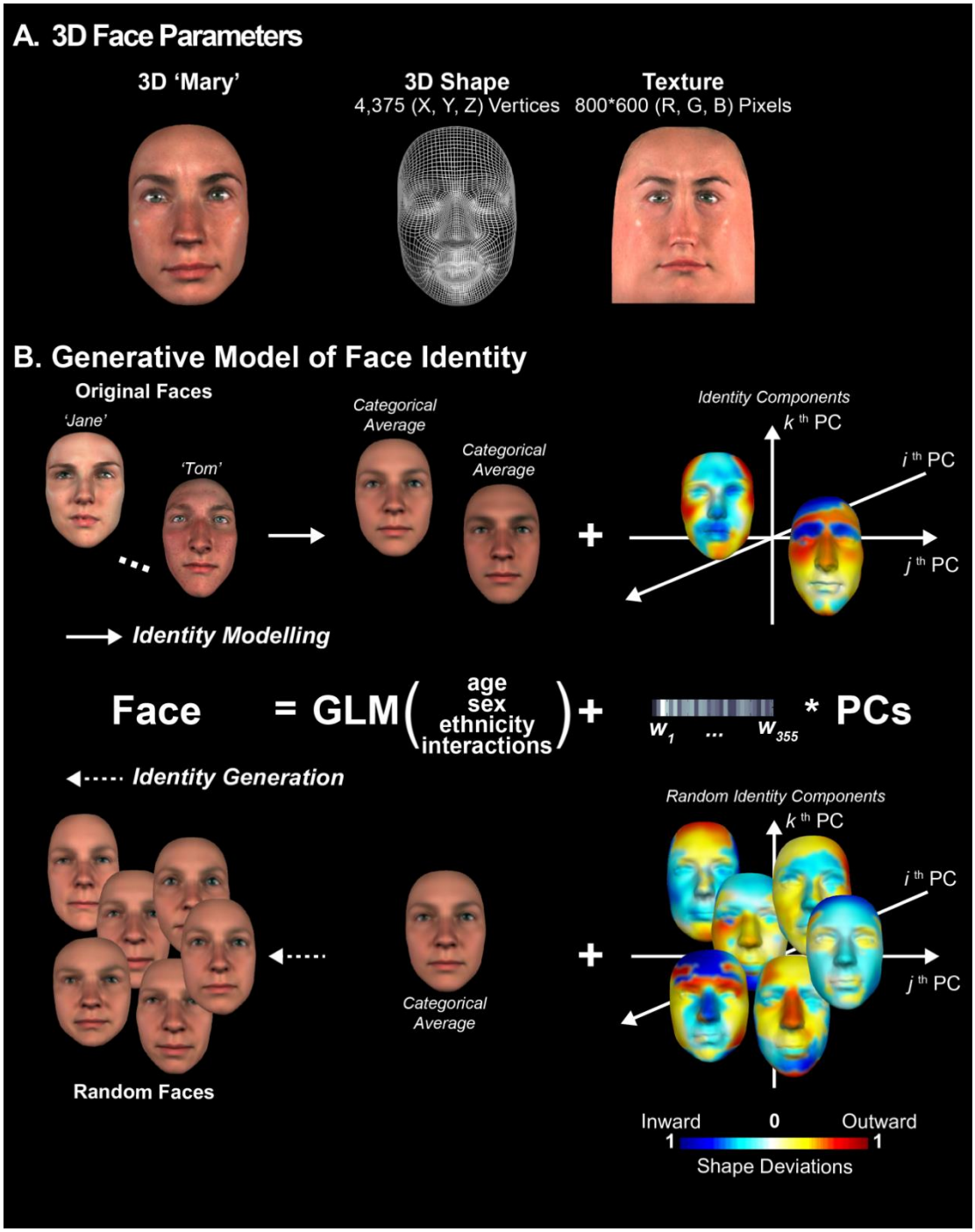

Figure 2–1 Generative Model of 3D Face Identity. (A) 3D Face parameters. We parameterized the shape of a face with the 3D coordinates of 4,735 vertices and its texture with 800*600 RGB 2D pixels. (B) Generative model of face identity. In its forward computation flow (see identity modelling solid arrow), the General Linear Model (GLM) decomposes a 3D, textured face (e.g. 'Jane' or 'Tom') into a non-identity face shape average capturing the categorical factors of face sex, ethnicity, age and their interactions plus a separate component that defines the identity of the face (illustrated by the 3D shape decomposition; 2D texture, not illustrated, is independently and similarly decomposed). Heat maps indicate the 3D shape deviations that define 'Jane' and 'Tom' in the GLM in relation to their categorical averages. In the reverse flow (see dashed arrow of identity generation), we can randomize the 3D shape identity component (and 2D texture component, not illustrated here), add the categorical average of 'Jane' (or 'Tom')

and generate random faces, each with a unique identity that share all other categorical face information with ‘Jane’ and ‘Tom.’

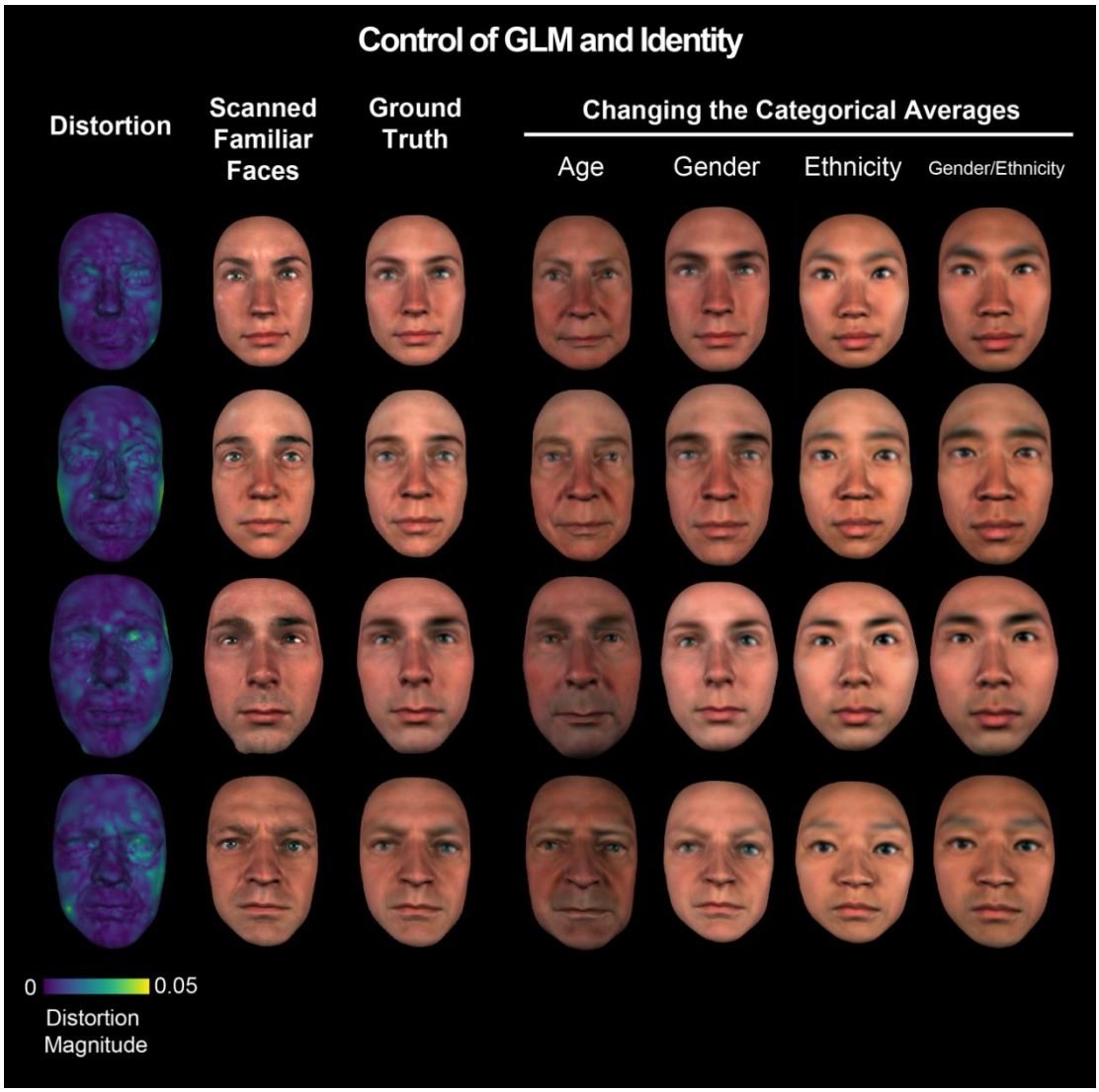

Figure 2–2 Control of Non-identity and Identity Information. *Distortion* quantifies, vertex per vertex, the quality of the 3D GLM fit of the scanned familiar faces, Color scale indicates the normalized Euclidean distance between the 3D positions of each vertex in the scanned face and the GLM fit. *Changing the categorical averages* illustrates, in each column, the GLM controls the factors of sex, ethnicity, and age using local averages, while the identity residuals are kept constant.

2.2.3 Stimuli

2.2.3.1 Four Familiar Faces

I scanned four faces ‘Mary’ and ‘Stephany’ (white Caucasian females of 36 and 38 of age, respectively), and ‘John’ and ‘Peter’ (white Caucasian males of 31 and 38 years of age, respectively) who were familiar to all participants as work colleagues (see Figure 2-2, 2nd column). These four familiar faces **were not** part of

the face database used to produce the 3D GMF introduced above. As I will explain, I used these scanned faces to compare the objective and mentally represented identity information in each participant.

2.2.3.2 Random Face Identities

For each of familiar face, I proceeded in three steps: First, I fitted the familiar identity in the GLM to isolate its non-identity averages, independently for shape and texture. Second, I randomized identity information by creating random identity residuals—i.e., I generated random coefficients (shape: 355; texture: 355*5) and multiplied them by the principal components of residual variance (shape: 355; texture: 355*5). Finally, I added the random identity residuals to the GLM averages to create a total of 10,800 random faces per familiar identity used in this experiment. Critically, each random face shared other categorical face information (i.e. sex, age and ethnicity) with the familiar face it is generated from.

2.2.4 Procedure

Each experimental block started with a centrally presented frontal view of a randomly chosen familiar face (henceforth, the target). On each trial of the block, participants viewed six simultaneously presented random identities of the target, created by the 3D GMF. They displayed in a 2 x 3 array on a black background, with faces subtending an average of 9.5° by 6.4° of visual angle. I instructed participants to respond on one of 6 buttons to choose the face that most resembled the target. The six faces remained on the screen until response. Another screen immediately followed instructing participants to rank the similarity of their choice to the target, using a 6-point Likert scale (‘1’ = not similar, ‘6’= highly similar) with corresponding response buttons. Following the response, a new trial began. The experiment comprised 1,800 trials per target, divided into 90 blocks of 20 trials each, run over several days, for a grand total of 7,200 trials per participant. Throughout, participants sat in a dimly lit room and used a chin rest to maintain a 76 cm viewing distance. I ran the experiment using the Psychtoolbox for MATLAB R2012a.

To resolve the task, participants must compare the randomly generated faces presented on each trial with their mental representation of the familiar target in full frontal view. Therefore, each face selected comprises a match to the participant’s mental representation of the target, which is estimated by the

similarity rating of that face (see Figure 2-3 for an example trial of target face 'Mary').

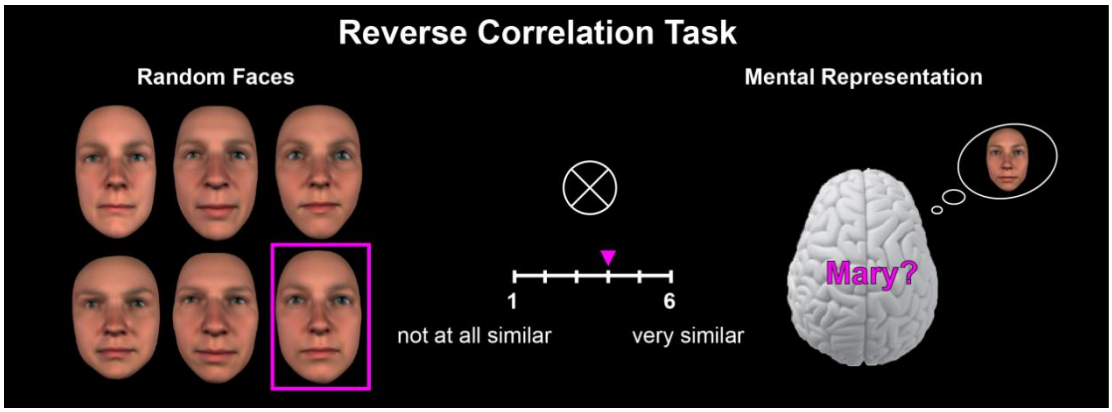

Figure 2–3 Illustrative Experimental Trial with 6 Randomly Generated Face Identities. I instructed participants to use their memory to select the face most similar to a familiar identity (here, 'Mary') and then to rate the similarity of the selected face (purple frame) to their memory of 'Mary' (purple pointer).

2.2.5 Analysis and Results

2.2.5.1 Linear Regression Model

For each participant and target face, each trial produced three outcomes: one matrix of 4,735*3 vertex parameters corresponding to the shape residuals of the chosen random face on this trial (c.f. face parameters in Figure 2-1A), one matrix of 800*600 pixels RGB corresponding to the texture residuals of this random face, and one corresponding rating response that captures the perceived similarity between the random face and the target.

To measure the relationship between random face shape parameters and the similarity rating, I proceeded in 3 steps:

Step 1: Across the 1,800 trials per target, I extracted the X, Y and Z coordinate value for one vertex and its corresponding similarity rating.

Step 2: I linearly regressed (i.e., RobustFit, Matlab 2013b) the coordinate value and similarity rating, separately for X, Y and Z. This linear regression produced a linear model with the 3D (i.e. X, Y, Z) Beta_1 and Beta_2 coefficients for this vertex.

Step 3: I repeated *Step 1* and *Step 2* for each of the 4735 3D vertices.

I applied the same analysis to measure the relationship between random face texture parameters and similarity rating, i.e. linearly regressing the RGB pixels (separately for R, G and B color channel) with the corresponding similarity rating values, and produced a linear model with 3D coefficients Beta_1 and Beta_2 for each RGB texture pixel.

Figure 2-4A schematizes the linear regression, using the 3D vertices and responses of Mary's trials from one participant. I plotted the 3D (i.e. X, Y, Z) Beta coefficients for each vertex on the heat maps by calculating their 3D Euclidean distance. The 3D Beta coefficients quantify the weighted changes (inward or outward) of each vertex in relation to the similarity rating in 3D face shape space.

We created for each participant and familiar face a linear model, separately for shape and texture. The linear models quantify how shape and texture identity residuals deviate from the GLM categorical average to represent the identity of each familiar face in the memory of each participant, which I used to reconstruct participant's memory representation (see the next section).

2.2.5.2 Reconstructing Mental Representations

Beta_2 coefficients can be amplified to control their relative presence in a newly synthesized 3D face. Figure 2-4B1 illustrates such amplification for one participant's Beta_2 coefficients of shape and texture of 'Mary.' Following the reverse correlation experiment, I brought each participant back to fine-tune their Beta_2 coefficients for each familiar face, using the identical display and viewing distance parameters as in the reverse correlation experiment.

In a self-adaptive procedure, I initialized Beta_2 amplification with equally spaced values between 0 and 50, with 10 unit increments. I then narrowed the amplification range to participant's responses until convergence, keeping the same total number of stimuli (i.e., 6 faces) per trial. Figure 2-4B2 illustrates the adaptive procedure.

The fine-tuning experiment comprised one session per familiar face, with familiar face order randomized across participants. Each session started with the screen presentation of the front view of one familiar face target to instruct participants as to the target of the session. On each trial, six faces initially amplified between 0 and 50 appeared on the screen, randomly positioned in a 2 by 3 array against a black background. I instructed participants to choose the face

that best resembled the familiar identity by pressing one of six response buttons. The six faces remained on the screen until response, immediately followed by the next trial. I repeated the trial five times, with the same six faces in different random array positions, to determine the next amplification range. I narrowed the amplification range every five trials by finding the minimum and maximum values that bound the participant's five choices. With this new range, I produced six new faces by evenly sampling the amplification values and again tested the participant over five new trials. I iteratively repeated sequences of five testing trials, updates of the amplification range, until it stabilized—i.e., remained constant over three blocks of five trials. I used the median of the final amplification range as value to generate the fine-tuned Beta_2 coefficients that I call *mental representation* in my analyses (see Figure 2-4B2).

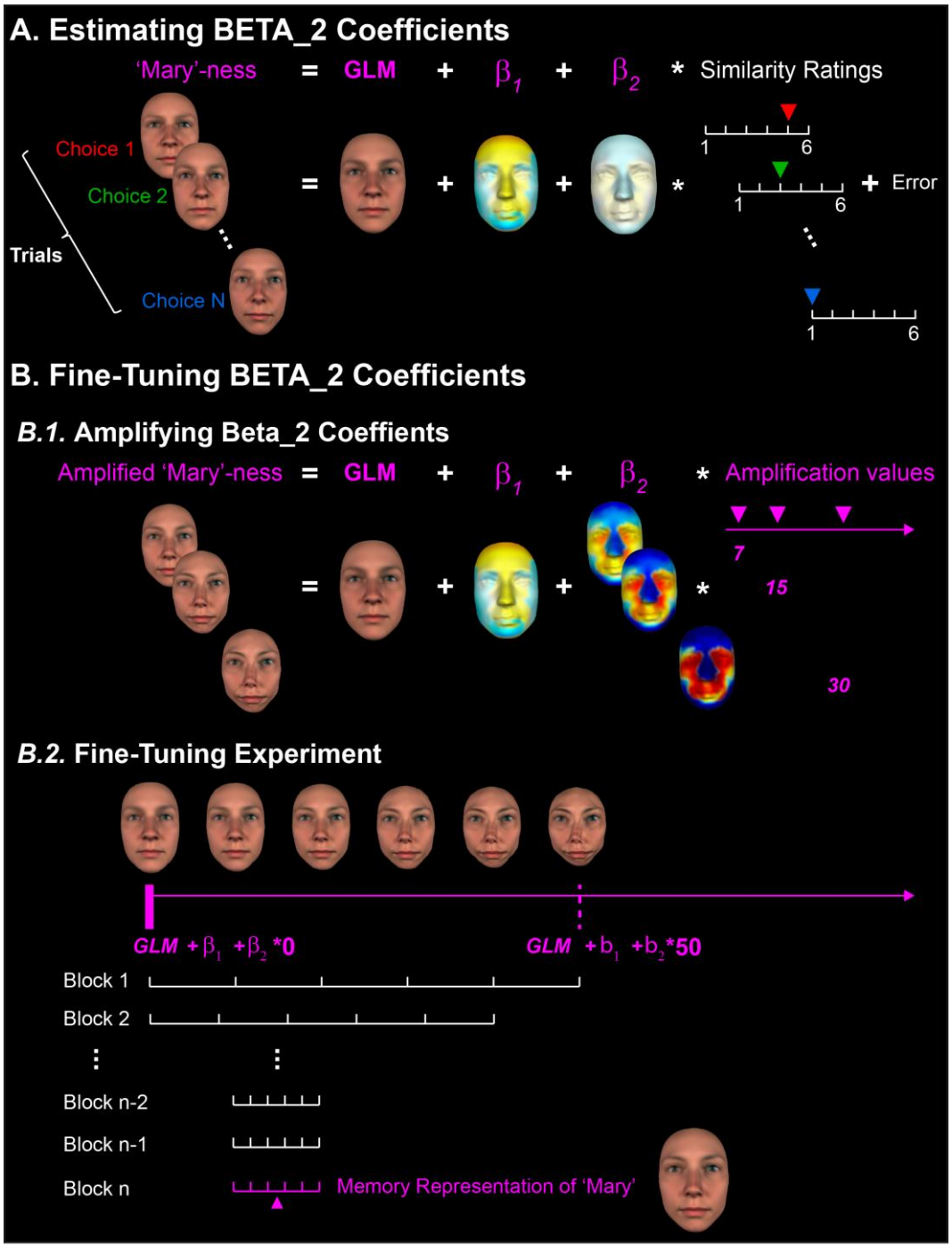

Figure 2–4 Reverse-Correlating the Information Contents of Familiar Face Representations. (A) Estimating Beta_2 Coefficients. I linearly regressed the 3D vertices of shape (separately for the X, Y and Z coordinates, texture not illustrated) with similarity judgments of the selected random identities (illustrated here for 'Mary'). For each vertex, 3D Beta_2 coefficients are color-coded according to their 3D magnitude (i.e. Euclidean Distance). Yellow-to-red indicates an outward change from the categorical average; turquoise-to-blue indicates an inward change from the categorical average. (B) Fine-tuning Beta_2 Coefficients. *(B.1) Amplifying Beta_2 coefficients.* Illustration of the amplification of Beta_2 coefficients. *(B.2) Illustration of the fine-tuning experiment.*

2.2.5.3 Vertex Contribution to Mental Representations

Vertices, whether in the ground truth face or in the participant's mental representation, can deviate inward or outward in 3D from the corresponding vertex in the common categorical average of their GLM fits (cf. Figure 2-1B). Thus, I can compare the respective deviations of their 3D vertices in relation to the common GLM categorical average.

For each participant and familiar face representation, I proceeded in three steps to classify each vertex as either 'faithful' or 'not faithful', and to test whether the vertices in mental representations deviated from the categorical average more than would be expected to occur by chance. I focus the analyses on the Beta_2 coefficients because they quantify how shape and texture identity residuals deviate from the GLM categorical average to represent the identity of each familiar face in the memory of each participant.

Step 1: I constructed a permutation distribution by iterating the regression analysis 1,000 times with random permutations of the choice response across the 1,800 trials. To control for multiple comparisons, I selected maximum (vs. minimum) Beta_2 coefficients across all shape vertices (and texture pixels), separately for the X, Y and Z coordinates (RGB color channels) from each iteration. I used the resulting distribution of maxima (and minima) to compute the 95% confidence interval of chance-level upper (and lower) Beta_2 value and classified each Beta_2 coefficient as significantly different from chance ($p < 0.05$), or not. I consider the vertex (or pixel) as significant if the Beta_2 coefficient of any coordinate (or color channel) was significant. There were very few significant pixels, with almost no consistency across participants (see Figure S2-1 in ***2.6.1*** *Supplemental Figures*), so I excluded texture identity residuals from further analyses.

Step 2: I used the chance-fit Beta coefficients in Step 1 and the Beta_2 amplification value derived in ***2.2.5.2 Reconstructing Mental Representation*** to compute the equation $GLM + \beta 1 + \beta 2 * amplification\ value$ (cf. Figure 2-4B2). As a result, I built a distribution of 1,000 chance fit faces.

Step 3: To classify whether each significant 3D vertex in the mental representation of a participant is more similar to ground truth than we would expect by chance, I computed D_{chance}, the mean Euclidean distance between the

1,000 chance fit faces and the veridical line, and D_{memory}, the distance between the same mental representation vertex and the veridical line. If $D_{memory} < D_{chance}$, this significant vertex is ***"Faithful"*** because it is significantly closer to the veridical line than chance. If $D_{memory} > D_{chance}$, and this vertex changes in the same direction as ground truth, the vertex is not faithful and I call it ***"Inaccurate+."*** If $D_{memory} > D_{chance}$, and this vertex changes in a direction opposite to the ground truth, the vertex is not faithful either, and I call it ***"Inaccurate-."***

Figure 2-5A shows results of one typical observer of familiar face 'Mary.' To illustrate, grey faces on the x-axis show the ground truth identity component in the GLM for Inward and Outward 3D shape deviations in relation to the categorical average (i.e., all white females of 30 years of age, like 'Mary'). For example, Mary's nose is objectively thinner than the average of white females of her age, and so these vertices deviate inward. Likewise, her more pouty mouth is shown as an outward 3D shape deviation. The y-axis of Figure 2-5A uses the same format to show the mental representation. The Inward and Outward orange-to-purple patches reveal faithful representations of, for example, a thinner nose and a pouty mouth. Blue patches denote 'inaccurate+' components of memory because they amplify (or caricature) the ground truth shape deviations. For example, a protruding part of the top lip exaggerates 'Mary's' pouty mouth. Green regions reveal 'inaccurate-' representation than changes in an opposite direction to the ground truth—i.e., inward residual vertices in memory when they are outward in the ground truth, or vice versa, such as the flatter surfaces between 'Mary's' nose.

A scatter plot visualizes the vertex by vertex fit between the mental representation (y-axis) and the ground truth 3D face (x-axis). The white diagonal line provides a veridical reference, where the identity component in the mental representation is identical to the ground truth face, for every single 3D vertex coordinates. As shown, the specific vertices near the veridical line faithfully represent 'Mary' in the mind of this participant ($p < 0.05$, two-tailed), as orange-to-red colored dots reported on the scatter and located on the y-axis faces in Figure 2-5A. In contrast, blue, green and the nonsignificant (white) vertices away from the veridical line did not faithfully represent the face.

I repeated the analysis of represented contents for each participant (N = 14) and each familiar face (N = 4). To derive group results, I counted across participants the frequency of each faithful/accurate+/accurate- vertex and used a

Winner-Take-All scheme to determine group-level consistency. For example, if 13/14 participants represented this particular vertex as ‘faithful,’ I categorized it as such at the group level. Figure 2-5B reports the collated group results, using the format of Figure 2-5A, where colors now indicate N, the number of participants who faithfully/ inaccurately+/inaccurately- represented that identity in their mind with this particular 3D shape vertex. Figure 2-5B demonstrates that mental representations comprised similar information contents across the 14 individual participants. Most (10/14) faithfully represented ‘Mary’s’ thin nose, ‘John’s’ receding eyes and wider upper face (13/14), ‘Peter’s’ prominent eyebrow and jawline (13/14), ‘Stephany’s’ protruding mouth (13/14).

To demonstrate the represented contents of individual participants comprise effective identity information, in next experiment, I tested these contents in a new set of participants.

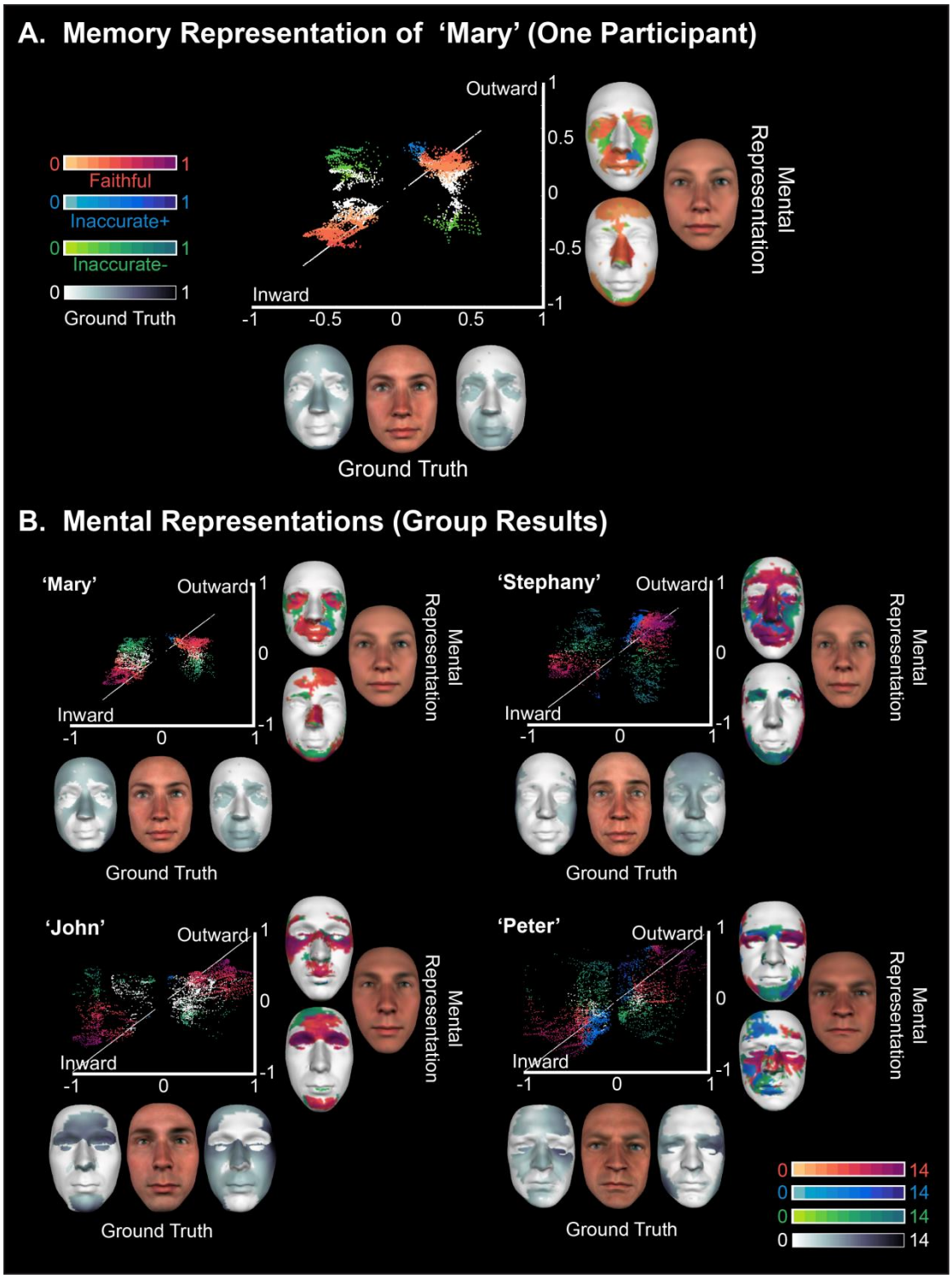

Figure 2–5 Contents of Mental Representations of Familiar Faces. (A) Mental representation of 'Mary' (a typical participant). *Ground truth*: 3D vertex positions deviate both Inward (-) and Outward (+) from the categorical average to objectively define the shape of each familiar face identity. Greyscale values reported on the flanking faces color-code the normalized magnitudes of inward and outward deviations from the categorical average. *Mental representation*: Inward and Outward red/blue/green face patches highlight the individual 3D vertices whose position faithfully/inaccurately+/inaccurately- deviate from the categorical average in the GLM. Color intensity represents the normalized magnitudes of their deviations. *2D scatter plots:* Scatter plots indicate the relationship between each vertex deviation in the ground truth (rank-ordered on a normalized scale on the X-

axis) and the corresponding vertex in the mental representation (also rank-ordered on a normalized scale on the Y-axis). The white diagonal line provides the reference of veridical mental representation in the GLM—i.e., a hypothetical numerical correspondence between each shape vertex position in the ground truth face and in the mental representation of the same face. White dots indicate vertices that were not significant in the linear regression. (B) Mental Representations (group results). Same caption as Figure 2-5A, except that the colormap now reflects the number of participants (N = 14) who faithfully/inaccurately+/inaccurately- represented this particular shape vertex.

2.3 Experiment 2: Validating the Contents of Mental Representations of Familiar Faces

2.3.1 Participants

I recruited 20 Western Caucasian participants (15 females and 5 males, mean age = 31.15 years, SD = 7.47 years). Each participant had to be familiar with 1 to 4 of the identities as work colleges and only participated in validation of familiar identities, up to 10 participants per identity. Table S2-2 in ***2.6.2*** *Supplemental Tables* reports the familiarity ratings of each identity for each participant.

All participants with normal or corrected-to-normal vision participated. As per self-report, no participant had history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (e.g., depression, autism spectrum disorder or prosopagnosia). All gave written informed consent and received £6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

2.3.2 Stimuli

I tested the 64 mental representation models (14 participants × 4 familiar faces) obtained in Experiment 1, together with 70 new random faces generated for each familiar identity (use the same way as described in *Experiment 1,* ***2.2.3.2.*** *Random Face Identities*).

2.3.3 Procedure

Participants ran a block of 14 trials for each familiar identity (i.e., the target) they were familiar with. In each identity block, we randomly allocated 14 models

and 70 random faces into 14 trials, with one trial displayed one model plus 5 random faces. Each trial started with the centrally displayed name of the target on a black background, which remained on the screen until participants pressed a button to start the trial. From the six faces randomly positioned in the 2 by 3 array on each trial, participants selected (with the mouse), the face most resembling the target, followed by the face least resembling the target. A 1.5s interval separated individual trials and a 30s break separated each target identity block. All viewing parameters were identical to the reverse correlation experiment.

Throughout, participants sat in a dimly lit room and used a chin rest to maintain the viewing parameters identical to Experiment 1. I ran the experiment using the Psychtoolbox for MATLAB R2012a.

2.3.4 Analysis & Results

We measured identification performance for each memory representation model as a percentage -- the number of participants who selected the model as the face most resembling the familiar target divided by the total number of participants (see Table 2-1).

Table 2-1. Recognition performance of mental representation models for the four familiar identities.

Model NO.	Mary	Stephany	John	Peter
1	1	0.8	1	1
2	1	0.8	1	1
3	1	0.7	1	0.9
4	1	0.7	1	0.9
5	1	0.7	1	0.9
6	0.9	0.6	0.9	0.9
7	0.7	0.6	0.9	0.9
8	0.7	0.6	0.8	0.9
9	0.7	0.5	0.8	0.8
10	0.7	0.5	0.7	0.8
11	0.6	0.4	0.7	0.8
12	0.6	0.2	0.6	0.8
13	0.5	0.1	0.5	0.8
14	0	0.1	0.4	0.5
Mean	0.74	0.52	0.81	0.85
SD	0.28	0.24	0.2	0.12

Note: the recognition performance of “Peter’s” models is saturated, with 13 out of 14 models of Peter showing excellent performance (>= 0.80, red outlined).

To understand the identification performance, I further tested the relationship (RobustFit, Matlab 2013b) between identification performance and the efficiency of mental representation derived by Equation (1), by pooling the data across 14 models of three identities. I withdrew the models of 'Peter' from this analysis because they saturated identification performance (red-framed in Table 2-1).

$$Efficiency = \frac{Number_{Faithful} - Number_{Inaccurate}}{Total_{Significant}} \qquad (1)$$

I found a robust positive correlation (r = 0.835, p < 0.001) between the efficiency index of individual models and their identification performance (see Figure 2-6).

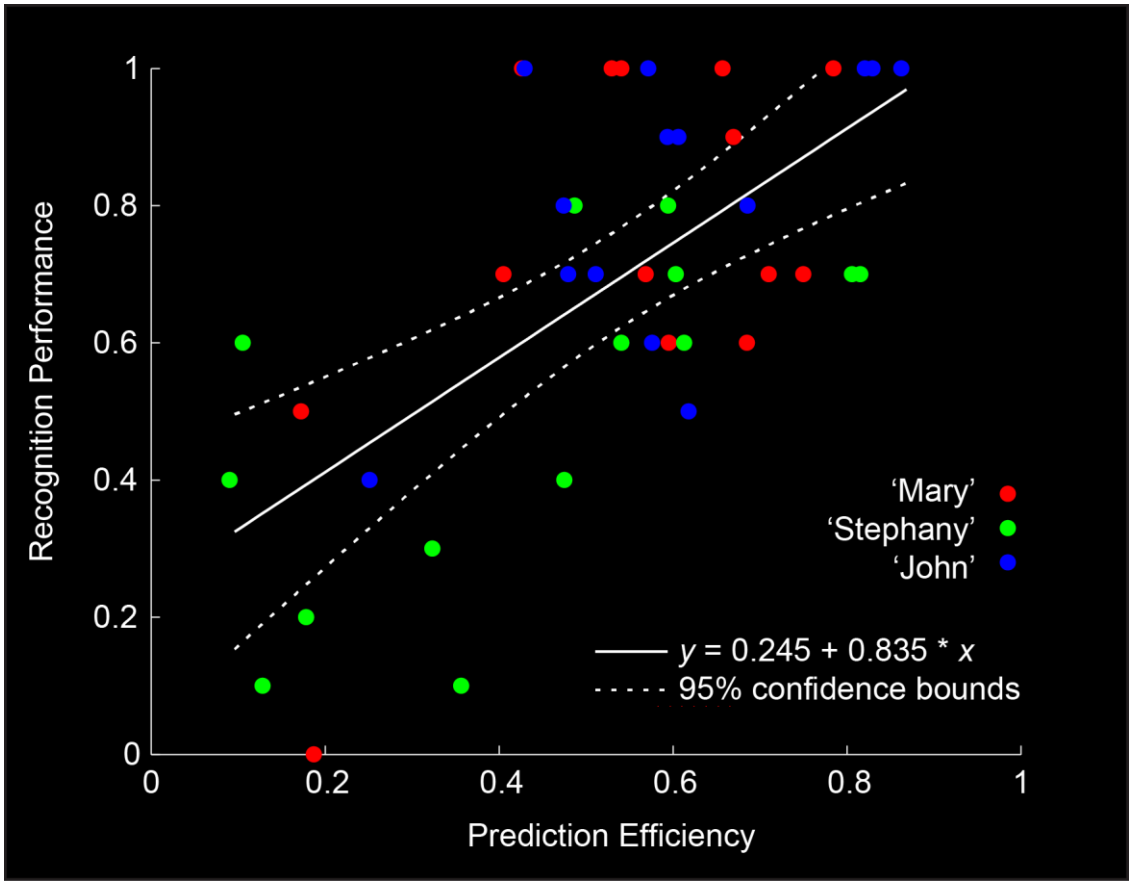

Figure 2–6 Model Efficiency and Recognition Performance. Scatter plots indicate the positive relationship between the model efficiency (X-axis) and the recognition performance of the models (Y-axis). The scattered points color-code recognition performance of the 14 models for 'Mary' (red), 'Stephany' (green) and 'John' (blue). I provide the robust fit together with the 95% confidence interval.

2.4 Experiment 3: Efficacy of the Information Contents of Mental Representations in New Participants and Tasks

Experiment 1 showed the convergence of represented contents across participants (cf. Figure 2-5), which suggests that the face representations could be multivariate (i.e., comprising contiguous surface patches rather than isolated vertices). Experiment 2 validated these models and shows that the faithful representational contents drive the face identification (cf. Figure 2-6). In this final experiment, I extracted the main multivariate components of faithfully represented surface patches, and validated their general use to other resemble tasks that preserve the identity identification – i.e., changes of viewpoints, age, and sex.

2.4.1 Participants

I recruited 11 Western Caucasians (7 females) and 1 East Asian (female), with mean age = 28.25 years and SD = 4.11 years. Each is familiar with the identities as work colleagues and also assessed familiarity on a 9-point Likert scale (see ***2.6.2*** *Supplemental Tables, Table S2-3*). All participants had normal or corrected-to-normal vision, without a self-reported history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (e.g., depression, autism spectrum disorder or prosopagnosia). They gave written informed consent and received £6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

2.4.2 Stimuli

2.4.2.1 Extracting Diagnostic vs. Nondiagnostic components of Mental Representations

To find common diagnostic components (multivariate features) that emerged in the group-level memory representation of each face identity, I factorized with Non-negative Matrix Factorization (NNMF, *D. D. Lee & Seung, 1999*) the total set of memory representations models across familiar identities and observers.

For each model, I recoded each vertex as ‘faithful’ = 1, ‘inaccurate+,’ ‘inaccurate-’ and not significant = 0, resulting in a 4735-d binary vector. I pooled 56 such binary vectors (across 4 targets x 14 observers = 56) to create a 4735 by 56 (i.e. vertex-by-model) binary matrix to which we applied NNMF to derive 8

multivariate components that captured the main features that faithfully represent familiar faces in memory across participants (Figure 2-7A shows each NNMF component).

To determine the loading (i.e., the contribution) of each NNMF component in the group-level mental representation of each familiar face identity, I computed the median loading of this component on the 14 binary vectors representing this identity in the 14 observers. I applied a 0.1 loading threshold (> 73 percentile of all 8 components × 4 identities median loadings) to ascribe a given component to a familiar face representation. The colored boxplot in Figure 2-7A represents the loading of each NNMF component at the group-level representation, showing that at least 2 above-threshold NNMF components represent each familiar identity.

I then constructed the diagnostic component of a familiar identity representation as follows: for each vertex we extracted the maximum loading value across the NNMF components representing it, and normalized the values to the maximum loading across all vertices. This produced a 4735-d vector V_d that weighs the respective contribution of each 3D vertex to the faithful representation of this familiar identity that we call the "diagnostic component." The heat maps in the left column of Figure 2-7B represent the diagnostic component of each familiar identity.

Crucially, I was then able to define a nondiagnostic component as the complement of the diagnostic component $V_n = 1 - V_d$, which capture variable face surfaces that do not comprise the participants' faithful mental representations. It is important to emphasize that I adjusted the total deviation magnitude of the diagnostic and nondiagnostic components from the categorical average—i.e., by equating the total sum of their deviations. Such normalization ensures that diagnostic and nondiagnostic components are both equidistant from the average face in the objective face space. The right column of Figure 2-7B shows the nondiagnostic component of each familiar identity representation.

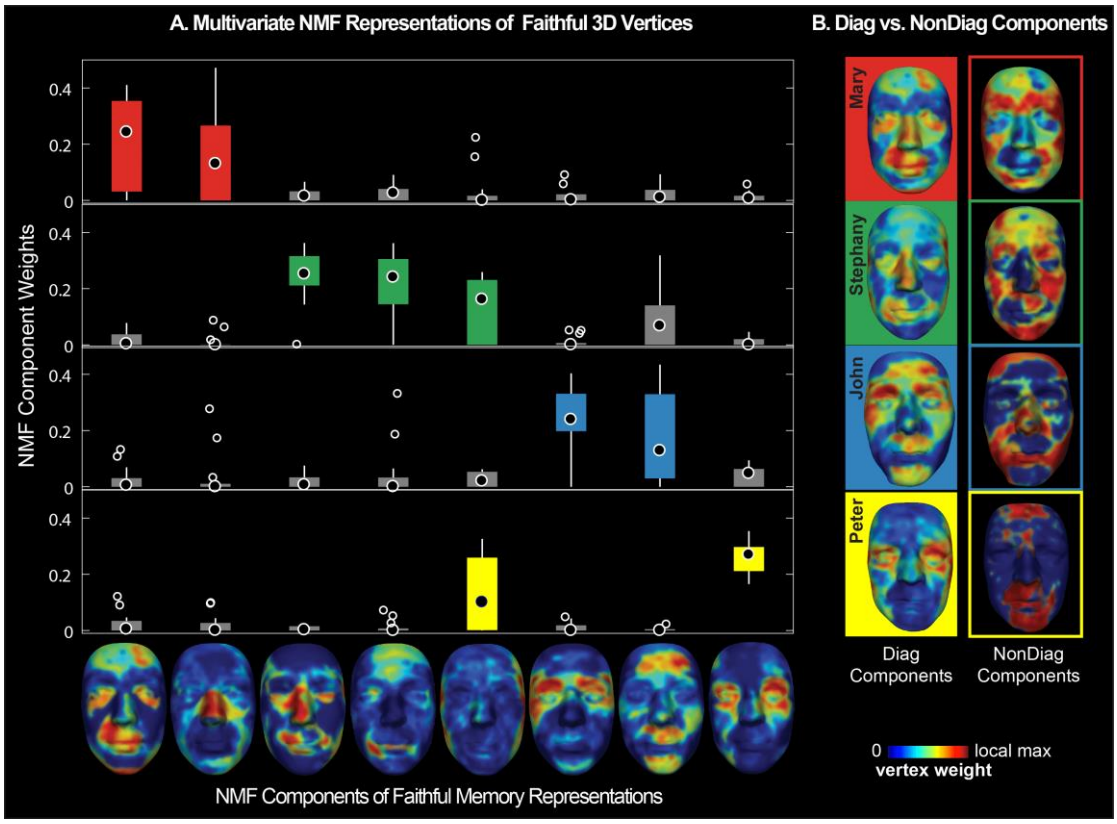

Figure 2–7 NNMF Multivariate and Compact Representations. A. NNMF representations of faithful 3D vertices across the mental representations of participants. The x-axis presents each NNMF component, where colors indicate the relative weight of each shape vertex in the component (normalized by maximum weight across components). Boxplots on the y-axis show the loading of each NNMF component on the faithful representations (N = 14, one per participant) of each familiar identity (N = 4 familiar identities), with colored boxes indicating above 0.1 threshold loading for NNMF components. B. Diagnostic and nondiagnostic components for each familiar identity. Heat maps in the left column show the group-level diagnostic component for each familiar identity; heat maps in the right column show the complementary nondiagnostic components.

2.4.2.2 Synthesizing Diagnostic and Nondiagnostic Faces

For each familiar identity, I synthesized new 3D faces that comprised graded levels of either the diagnostic or the nondiagnostic shape components. Specifically, I used the normalized diagnostic component V_d and its nondiagnostic complement V_n to synthesize morphed faces with shape information of each target identity as follows:

$$Diagnostic\ Faces = Ground\ Truth\ \times Vd \times \alpha + Categorical\ Average\ (1 - Vd \times \alpha)$$

$$Nondiagnostic\ Faces = Ground\ Truth\ \times Vn \times \alpha + Categorical\ Average\ (1 - Vn \times \alpha)$$

with amplification factor α = 0.33, 0.67, 1, 1.33, 1.67, to control the relative intensity of diagnostic and nondiagnostic shape changes. I rendered all these morphed shapes with the same average texture.

I also changed the viewpoint, age, and sex of all of these synthesized faces (cf. *Experiment1, Generative Model of Face Identity*). Specifically, I rotated them in depth by -30 deg, 0 deg and +30 deg and using the 3D GMF; I set the age factor to 80 years/swapped the sex factor, keeping all other factors constant. Figure 2-8A shows the diagnostic and nondiagnostic faces of 'Mary' with increasing amplification values and any changes of viewpoints/age/sex (see ***2.6.3*** *Supplemental Figures,* Figure S2-2 to Figure S2-5 for all familiar faces). I added as filler stimuli the grand average face (for both shape and texture) of the 355 database faces.

It is important to emphasize that both diagnostic and nondiagnostic faces are equally faithful representations of the original ground truth. That is, their shape features are equidistant from the shared categorical average. However, whereas the diagnostic components deviate from the average with multivariate information extracted from the participants' mental representations, the nondiagnostic components do not. I hypothesized that, though equidistant from the categorical average, only the diagnostic components will effectively impact performance on the novel tasks.

2.4.3 Procedure

The experiment comprised 3 conditions (viewpoint, age and sex) that all validators accomplished in a random order, with one condition per day. In the Viewpoint condition, validators ran 15 blocks of 41 trials (5 repetitions of 123 stimuli). Each trial started with a centrally displayed fixation for 1s, followed by a face on a black background for 500ms. I instructed participants to name the face as 'Mary,' 'Stephany,' 'John' or 'Peter,' or respond 'other' if they could not identify the face. Participants were required to respond as accurately and as quickly as possible. A 2s fixation separated each trial. Validators could break between blocks. In the Age and Sex conditions, validators ran 5 blocks that repeated 42 trials. They were instructed to respond "Old 'Mary," "Old Stephany," "Old John," "Old Peter" or "Other" in the age condition, and "Mary's brother", "Stephany's brother," "John's sister," "Peter's sister" or "Other" in the sex condition.

2.4.4 Analysis & Results

For each participant and generalization condition, I computed the percent correct identification of diagnostic and nondiagnostic faces for each familiar face and at each level of feature intensity. In each of the three Viewpoints, I paired the performance of diagnostic and nondiagnostic stimuli of each participant and identity, at each level of feature intensity (i.e., 5 levels x 4 identities x 12 participants = 240 pairs). In Age and Sex, I paired the performance of diagnostic and non-diagnostic stimuli of each validator and identity at each intensity level (i.e., 5 levels x 4 identities x 12 validators = 240 pairs). Then, in each of the five tasks (3 views plus age and gender), I used a Wilcoxson sign ranked test to compare the diagnostic vs. nondiagnostic pairs and obtained the Wilcoxson sign ranked statistics *W*. I determined the statistical significance of *W* using a bootstrapped null distribution of *W* (N = 1,000 iterations). On each iteration, I randomly shuffled the participants' responses across stimuli within a task, and computed the *W* between two random pairs. I use the 99th percentile of resulting distribution as the statistical threshold (i.e., $p < 0.01$, one-tailed). I corrected significance across the five tasks—i.e., Bonferroni corrected $p < 0.05$, 1-tailed.

In each task, I found a significantly higher global identification performance for diagnostic faces (see Figure 2-8B, red curves) than for nondiagnostic faces (black curves, $p < 0.05$, 1-tailed. Thus, the diagnostic contents of the memory representations I modelled do indeed contain the information that can resolve identity and novel resemblance tasks.

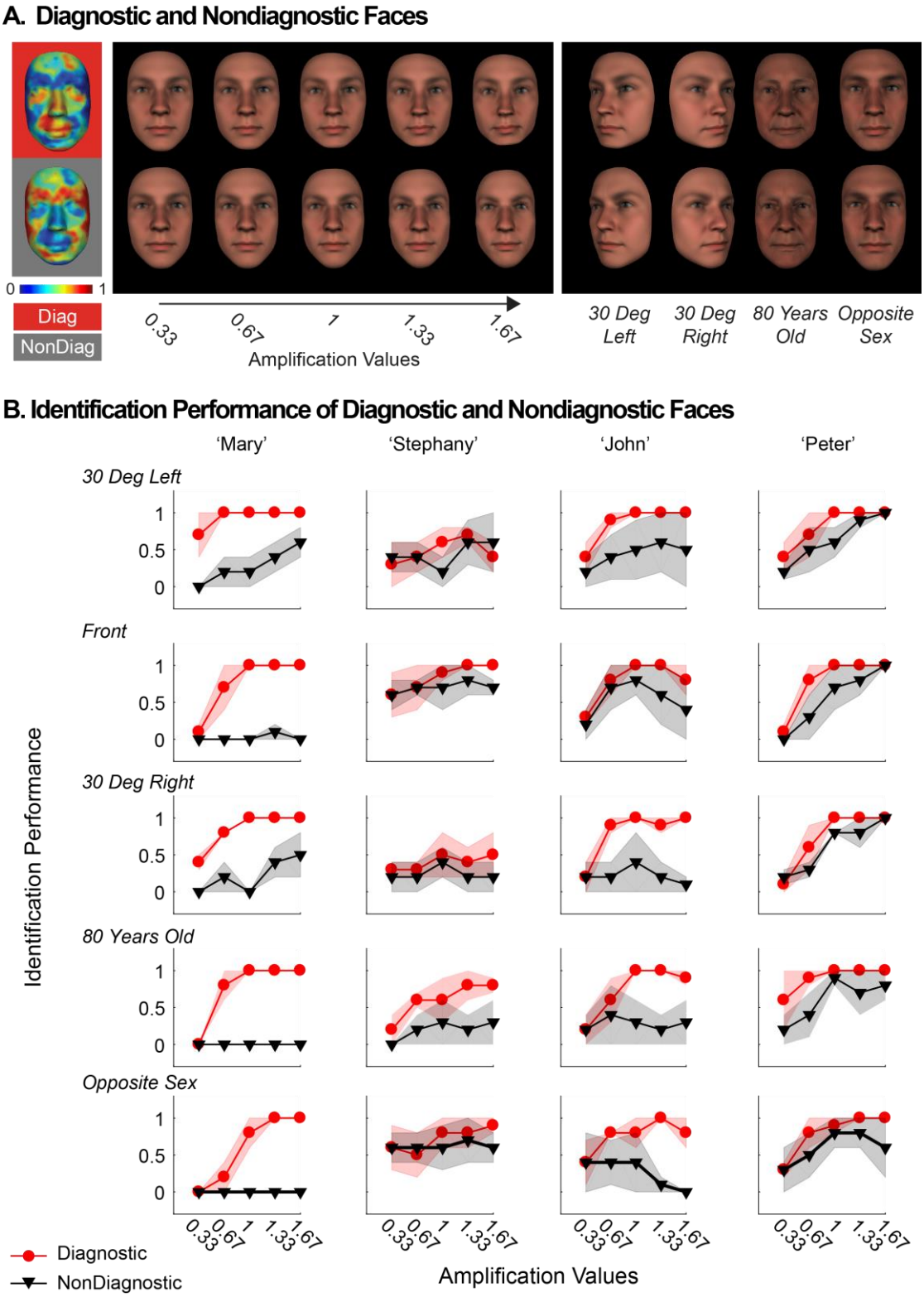

Figure 2–8 Generalization of Performance across Tasks. (A) Diagnostic and nondiagnostic Faces. *Left panel:* The red framed map shows the multivariate diagnostic components of faithful 3D shape representation of 'Mary'; the grey framed map shows the nondiagnostic complement (1 - diagnostic components). *Middle panel:* Faces synthesized with increasing amplification (0.33 to 1.67) of the diagnostic (top) vs. nondiagnostic (bottom) components. *Right panel:* For each synthesized face, we changed its viewpoint (30° left and 30° right), age (80 years old) and sex, shown here for faces synthesized at amplification = 1. (B) Task Performance. For each condition of generalization (y-axis) and familiar identity (x-axis), 2D plots show the median identification performance computed across

participants (y-axis) for faces synthesized with the diagnostic (red curves) and nondiagnostic (grey curves) faces, at different levels of amplification of the multivariate components (x-axis). Shadowed regions indicate median absolute deviations (MAD) of identification performance.

2.5 Discussion

Mental representations stored in memory are critical to guide the information processing mechanisms of cognition. Here, with a novel methodology based on reverse correlation and a new 3D face information generator, I modelled the information contents of mental representations of 4 familiar faces in 14 individual participants. I showed that the contents converged across participants on a set of multivariate features (i.e., local and global surface patches), which faithfully represent 3D information that is objectively diagnostic of each familiar face. Critically, I showed that validators could identify new faces generated with these diagnostic representations across new resemblance tasks—i.e., changes of pose, age and sex—but performed much worse with equally faithful, but nondiagnostic features. Together, these results demonstrate that the modelled representational contents were both sufficiently precise to enable face identification within task and versatile enough to generalize the face identification to other resemblance tasks.

There has been a recent surge of interest in modelling face representations from human memory *(C. H. Chang et al., 2017; H. Lee & Kuhl, 2016; Nestor et al., 2016*). These studies used 2D face images and applied dimensionality reduction (e.g. PCA ,*Turk & Pentland, 1991;* and multidimensional scaling) to formalize an image-based face space, where each dimension is a 2D eigenface or classification image – i.e. pixel-wised RGB (or L*A*B) values. To understand the contribution of each 2D face space dimension to memory representations (including their neural coding), researchers modelled the relationship between projected weights of the original 2D face images on each dimension and participants' corresponding behavioral *(C. H. Chang et al., 2017*) and brain *(H. Lee & Kuhl, 2016; Nestor et al., 2016*)) responses.

These studies contributed important developments in face identification research because they addressed the face identity contents that the brain uses to guide face identification mechanisms. The aim of my experiments was to model the face identity contents in the generative 3D space of faces (not the 2D space of their image projections) and to use these models to generate identification

information in resemblance tasks that test the generalizability of identity information. It is important to clarify that I modelled identity information in a face space that belongs to the broad class of 3D morphable, Active Appearance Models of facial synthesis (AAMs, *Blanz & Vetter, 1999; Cootes, Edwards, & Taylor, 2001*). These models contain full 3D surface and 2D texture information about faces and so with their better control superseded the former generation of 2D image-based face spaces (*Rhodes & Jeffery, 2006; Turk & Pentland, 1991; O'Toole, Castillo, Parde, Hill, & Chellappa, 2018*). To synthesize faces, I used our GMF to decompose each face identity as a linear combination of components of 3D shape and 2D texture added to a local average (that summarizes the categorical factor of age, gender, ethnicity and their interactions, cf. Figure 2-1B). To model the mental representations of faces, I estimated the identity components of shape and texture from the memory of each observer. These components had generative capacity and we used them to precisely control the magnitude of identity information in new faces synthesized to demonstrate generalization across pose, age and sex. Thus, I used the same AAM framework for stimulus synthesis, mental representation estimation and generation of generalizable identities.

There is a well-known problem with using AAMs to model the psychology of face recognition. Perceptual expertise and familiarity are thought to involve representations of faces that enable the greater generalization performance that is widely reported (*Eger, Schweinberger, Dolan, & Henson, 2005; Jenkins, White, Van Montfort, & Burton, 2011; White, Phillips, Hahn, Hill, & O'Toole, 2015; Young & Burton, 2018*). However, AAMs typically adopt a brute force approach to identity representation: a veridical (i.e., totally faithful) deviation of each physical shape vertex and texture pixel from an average. Thus, as AAMs overfit identity information, they appear as a priori weak candidate models to represent the perceptual expertise (*O'Toole et al., 2018*). My approach to studying the contents of mental representations offers a solution to this information processing conundrum. I showed that each observer faithfully represented only a proportion of the objective identity information that defines a familiar face identity. Its key theoretical contribution to face space is to formalize the diagnostic information as a reduced set of multivariate face features that can be construed as dimensions of the observer's face space. Observers develop these dimensions whenever they interact with the objective information that represents a new face identity in the real world. I modelled the objective information that is available to the observer for

developing their face space dimensions via learning as the veridical shape and texture information of the AAM (*Gosselin & Schyns, 2002; O'Toole et al., 2018; Schyns, 1998*). Key to demonstrating the psychological relevance of our psychological face space dimensions is that they should comprise identity information sufficiently detailed to enable accurate face identification and sufficiently versatile to enable similarity judgments of identity in novel tasks. My results demonstrated this potential when validators identified faces synthesized with the diagnostic dimensions in novel resemblance tasks. Thus, by introducing reduced faithful mental representations of identity information in the objective representations of AAMs my study provides the means of modelling the psychological dimensions of face space.

The practical contribution of my study to face recognition is that we can now precisely track the development of the psychological dimensions of face space. AAMs enable a tight control of objective face information at synthesis, such as ambient factors of illumination, pose and scale, but also categorical factors of gender, sex, age and ethnicity and components of identity. Thus, it is now possible to tightly control the statistics of exposure to faces in individual observers, model the diagnostic dimensions of the psychological face space that are learned, and finally test the efficacy of the psychological face space as I did here. And when we understand how ambient and categorical factors influence performance, we can switch to understanding familiar face identification in the wild, where all ambient and categorical factors get naturally mixed up, and where the influence of each factor to identification performance becomes near impossible to disentangle, precluding a detailed information processing understanding of face identification mechanisms.

Our results could suggest that the representation of face shape information trumps its texture. At this stage, it is important to clarify that shape and texture have different meanings in different literatures. For example, some authors in psychology discuss *shape-free faces* when referring to 2D images synthesized by warping an identity-specific texture to an identical ‘face shape’ (defined as a unique and standard set of 2D coordinates that locate a few face features (*Burton, Schweinberger, Jenkins, & Kaufmann, 2015)*). However, it is important to emphasize that the warped textures are not free of 3D shape information (e.g. that which can be extracted from shading, *Erens, Kappers, & Koenderink, 1993*). In

computer graphics, the generative model of a face comprises a 3D shape per identity (here, specified with 4,735 3D vertex coordinates), lighting sources (here, N = 4), and a shading model (here, Phong shading, *Phong, 1975*). The shading model interacts with shape and texture to render the 3D face as a 2D image. To illustrate the effects of this rendering, Supplementary Figure 9 shows how applying the same 2D textures (rows) to different 3D face shapes (columns) generates 2D images with different identities. We used the better control afforded by computer graphics to generate our face images and found that shaded familiar face shape was more prevalent in the face memory of individual participants than face texture.

The models of mental representation should be construed as the abstract information goals that the visual system predicts when identifying familiar faces. I term the faithful components as the 'abstract information goal' because it has to be broken down into global and local constituents according to the biological constraints of representation and implementation at each level of the visual hierarchy—or their analogues in a multi-layered deep convolutional network, where we can use a similar methodology to understand the identity contents represented in the hidden layers (Xu et al., 2018). In norm-based coding (*Leopold et al., 2001; Rhodes & Jeffery, 2006*), face identity information is represented in reference to the average of a multi-dimensional face space. Monkey single cell responses increase their firing rate with increasing distance of a face to this average (as happens with e.g. caricaturing, *Leopold et al., 2006*). As shown by Chang and Tsao (*2017*), neurons selectively respond along a single axis of the face space, not to other, orthogonal axes. An interesting direction of research is to determine whether our reduced diagnostic features, as defined by our 'abstract information goal' (see also *Zhan, Ince, van Rijsbergen, & Schyns, 2019*), provide a superior fit to the neural data than the full feature sets used in the axis model used by Chang et al. (*L. Chang & Tsao, 2017*).

Though I modelled the mental representation of a face identity in an AAM, it is important to state that I do *not* assume that memory really represents faces in this way (i.e., as demarcations to an average, separately for 3D shape and 2D texture). AAM is only a state-of-the-art, mathematical modelling framework. I fully acknowledge there are many possible concrete implementations into a neural, or a neurally-inspired architecture that could deliver AAM-like performance without assuming an explicit AAM representation. What is clear is that whichever

implementation, in whichever architecture, the abstract information represented will have to enable the performance characteristics my resemblance tasks demonstrated.

2.6 Supplemental Materials

2.6.1 Supplemental Methods

3D Face Database. The face database comprised 197 females, 158 males, 233 Western Caucasian, 122 East Asian, age between 16 and 86, SD = 15.06, scanned in-house with a Di4D face capture system, at a high resolution in shape (4,735 3D vertex coordinates) and texture (800*600 RGB pixels, see Figure 2-1A). All 3D models were in full color with hair removed, posing with a neutral facial expression.

2.6.2 Supplemental Tables

Table S 2-1 Identity familiarity ratings of 14 participants in Experiment 1.

Participants	Mary	Stephany	John	Peter
1	6	4	3	7
2	7	8	9	6
3	7	9	7	3
4	8	8	7	8
5	3	4	4	4
6	3	3	4	5
7	5	6	9	4
8	5	6	9	5
9	8	7	7	4
10	8	6	8	10
11	7	7	8	7
12	7	6	6	7
13	5	3	9	7
14	9	9	9	9
Mean	6.29	6.14	7.07	6.14
SD	1.86	2.03	2.09	2.07

Note: Ratings are from 1 (not familiar at all) to 9 (highly familiar).

Table S 2-2 Identity Familiarity rating of 20 participants in Experiment 2.

Validators	Mary	Stephany	John	Peter
1	6	7	7	5
2	9	9	9	8
3	5	--	4	6
4	5	--	5	7
5	7	--	7	7
6	9	9	--	9
7	6	--	5	--
8	7	--	--	6
9	4	8	--	--
10	5	--	--	--
11	--	9	--	7
12	--	5	--	--
13	--	6	--	--
14	--	5	--	--
15	--	7	--	--
16	--	5	--	--
17	--	--	5	7
18	--	--	6	8
19	--	--	6	--
20	--	--	7	--
Mean	6.3	7	6.1	7
SD	1.7	1.7	1.45	1.15

Note: Ratings are from 1 (not familiar at all) to 9 (highly familiar). "--" indicates validators did not give rating or participate in the experimental block, since they were not familiar with the target identities.

Table S 2-3 Identity familiarity ratings of 12 participants in Experiment 3.

Model NO.	Mary	Stephany	John	Peter
1	1	0.8	1	1
2	1	0.8	1	1
3	1	0.7	1	0.9
4	1	0.7	1	0.9
5	1	0.7	1	0.9
6	0.9	0.6	0.9	0.9
7	0.7	0.6	0.9	0.9
8	0.7	0.6	0.8	0.9
9	0.7	0.5	0.8	0.8
10	0.7	0.5	0.7	0.8
11	0.6	0.4	0.7	0.8
12	0.6	0.2	0.6	0.8
13	0.5	0.1	0.5	0.8
14	0	0.1	0.4	0.5
Mean	0.74	0.52	0.81	0.85
SD	0.28	0.24	0.2	0.12

Note: the recognition performance of “Peter’s” models is saturated, with 13 out of 14 models of Peter showing excellent performance (>= 0.80, red outlined).

2.6.3 Supplemental Figures

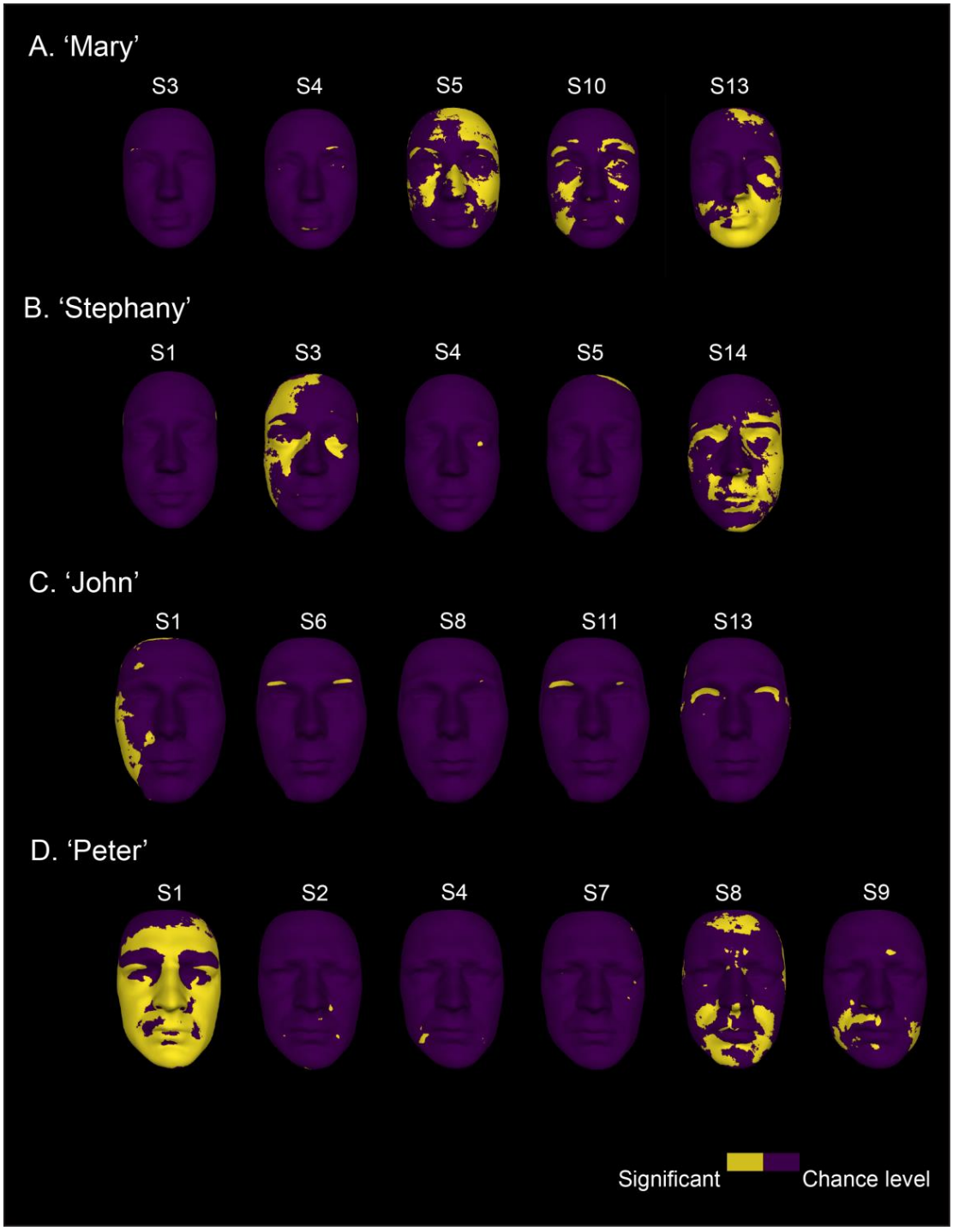

Figure S 2-1 Beta_2 Coefficients of Face Texture. Yellow colored overlays on each familiar face illustrate the significant Beta_2 coefficients for RGB texture pixels in each participant (labelled S1-S14). Dark purple pixels represent non-significant RGB coefficients.

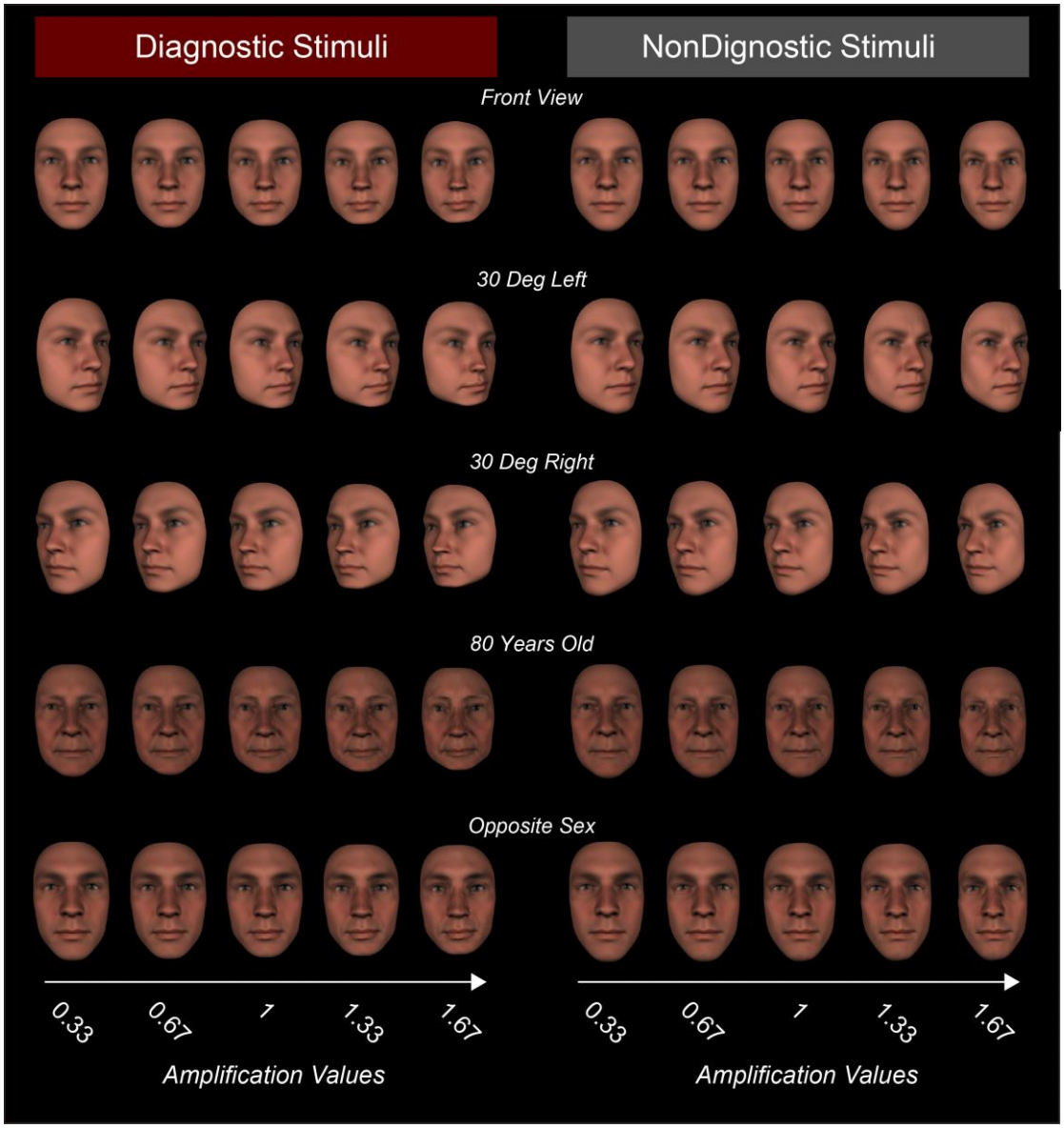

Figure S 2-2 Diagnostic (Left) and Nondiagnostic (Right) Faces of 'Mary.' Each row presents the main conditions of stimulus synthesis (i.e., 3 viewpoints, age and sex). Each column presents a level of diagnostic (vs. nondiagnostic) component amplification in the face.

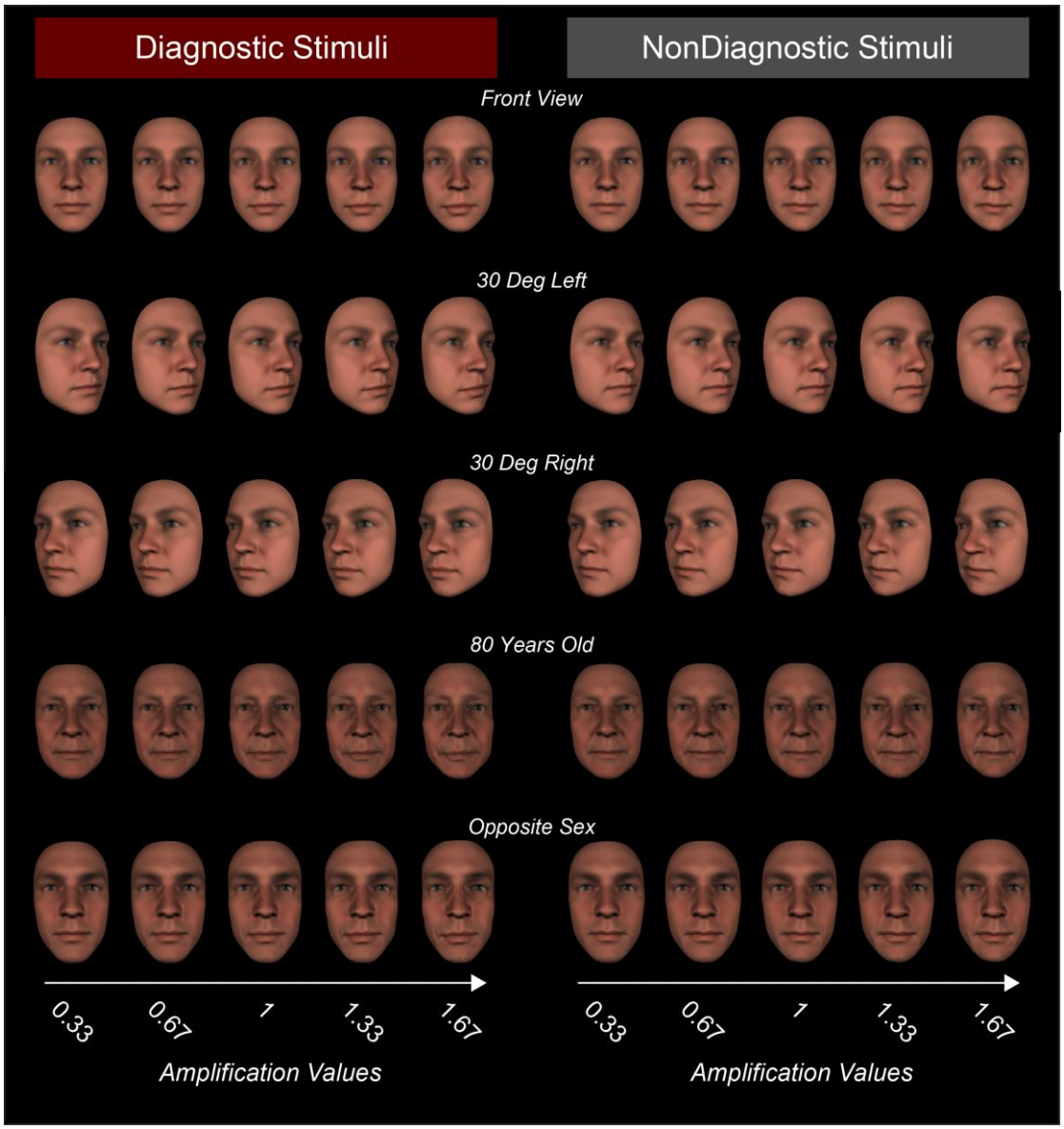

Figure S 2-3 Diagnostic (Left) and Nondiagnostic (Right) Faces of 'Stephany.' Same caption as in Figure S2-2.

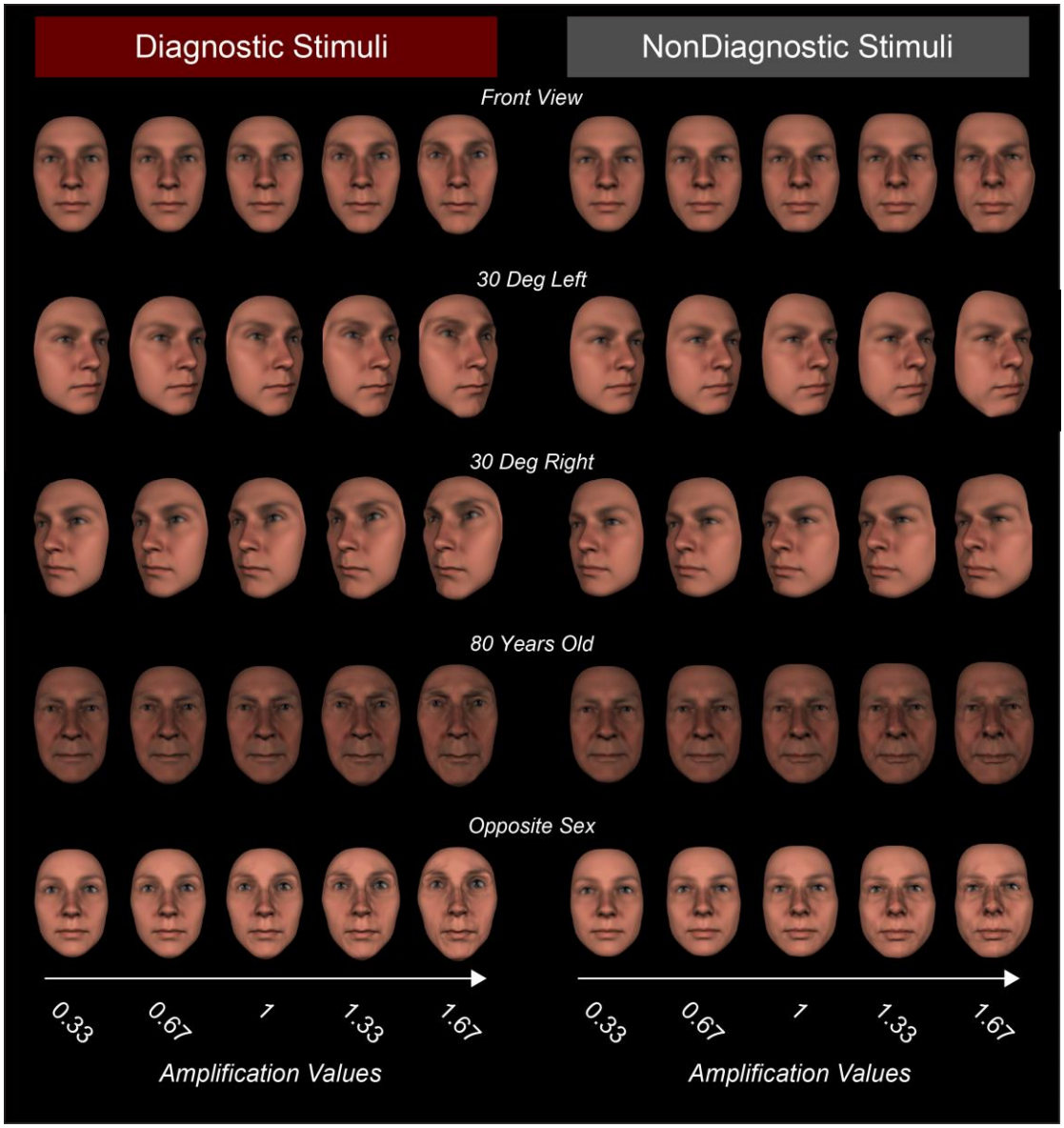

Figure S 2-4. Diagnostic (Left) and Nondiagnostic (Right) Faces of 'John.' Same caption as in Figure S2-2.

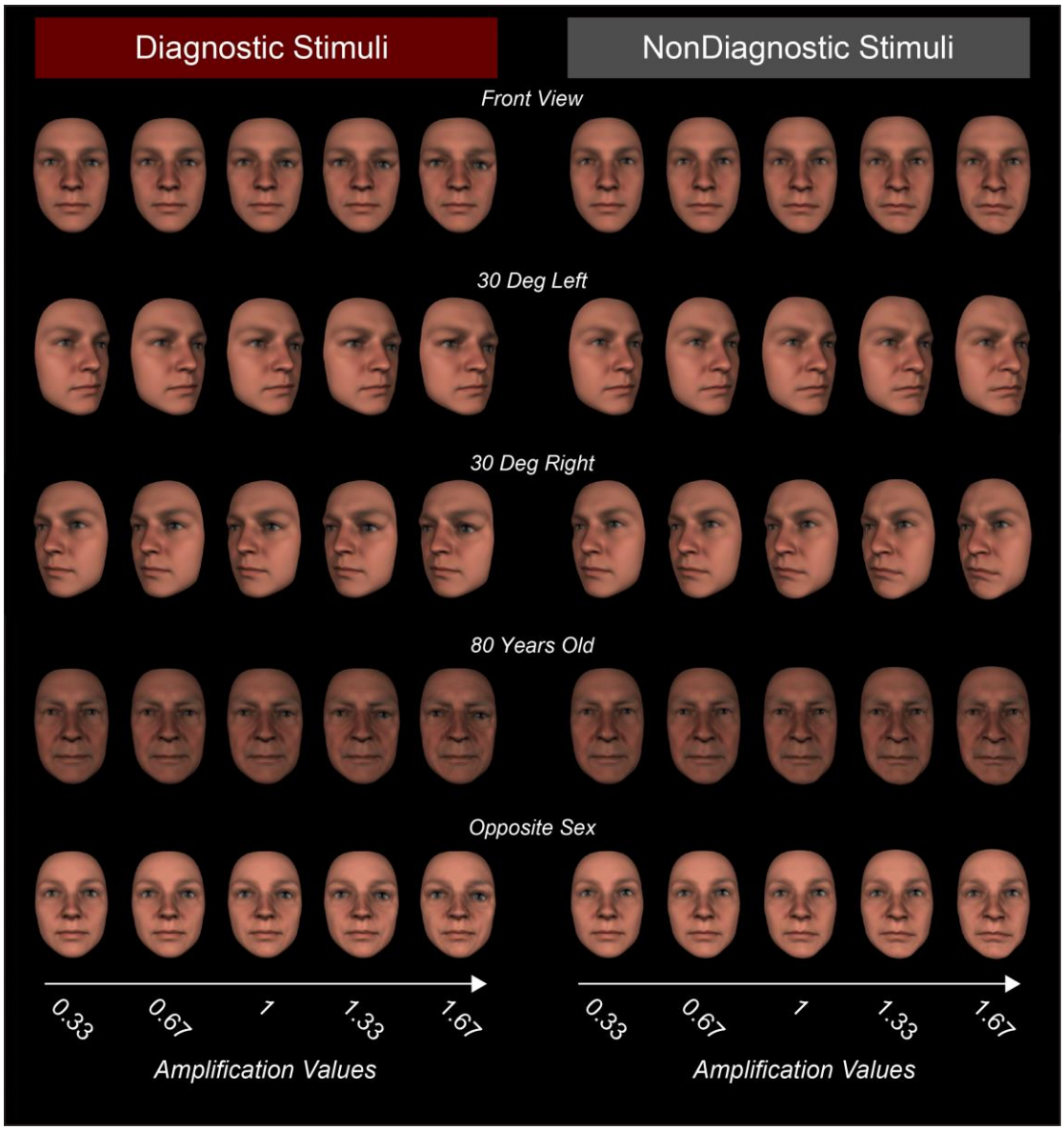

Figure S 2-5. Diagnostic (Left) and Nondiagnostic (Right) Faces of 'Peter.' Same caption as in Figure S2-2.

3 Study 2: Modelling the Diagnostic Information for Facial Expression Recognition

3.1 Introduction

Accurately recognizing facial expressions of emotion requires shared representation of the expressions in the mind between signal sender and receiver (*Jack & Schyns, 2015*). Such shared knowledge is perceived, consolidated and retained during individuals' interaction with the external environment *(Yuille & Kersten, 2006*), and is therefore expressed in a cultural-specific manner *(Blais, Jack, Scheepers, Fiset, & Caldara, 2008; Jack et al., 2009; Jack, Caldara, et al., 2012; Jack, Garrod, et al., 2012)*.

In this study, I used the facial expression models of six emotions (i.e. Happy, Surprise, Fear, Disgust, Anger and Sad) derived from Western observers in a reverse correlation experiment *(Jack et al., 2014)* and tested them in a new group of Western validators. Among these models, I selected a subset that shows good recognition performance as the validated models and use these models to derivd the main vairants of each emotion. Using these emotion variants and their probabilities of occurrence in the validated set, I trained a Bayesian classifier that showed a level of categorization performance mimics human observers closely. I explained these emotion variants together with their occurrence probability as the mental representation of six facial emotions for Western observers at the population level.

3.2 Experiment

3.2.1 Participants

Three hundred (150 female, 284 European, 15 North American, 1 Australian) Western white Caucasian (WC) participants (mean age of 20.9 years, SD = 2.948 years; range 18-30 years) with normal, or corrected to normal vision participated in the experiment. All participants had minimal experience of non-Western cultures (as assessed by questionnaire, see ***3.4*** *Supplemental Materials, Screening Questionnaire*), normal or corrected to normal vision, gave written informed consent and received £6 per hour. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval. Henceforth, we called these participants 'validators.'

3.2.2 Stimuli

I included 720 dynamic models (120 for each emotion) reconstructed from a reverse correlation experiment (Jack et al., 2014). I synthesized each model to a set of 50 WC faces (25 females and 25 males, between 18 and 31 years of age, mean = 22.5 years, SD = 3.69 years) obtained using standard face capture (*Yu, Garrod, & Schyns, 2012*). This produced a total of 36,000 dynamic validation stimuli (720 models x 50 identities).

To better understand these dynamic models and the analysis in *section 3.2.4*, it is necessary I introduce three crucial components in the reverse correlation experiment:

Generative Face Grammar (GFG). In the Generative Face Grammar, the generative model of facial expression signal is the dynamic 3D face, equipped with 42 independent and biologically plausible facial movements (called Action Unites, (Ekman & Friesen, 1978)). A subset of active Action Units (AUs) comprises a pattern of facial movements on 3D faces; and the temporal dynamics of each active AU is modelled using 6 parameters: onset, acceleration, peak amplitude, peak latency, deceleration and offset. GFG was designed to synthesize 3D face dynamics by tightly controlling the active AUs (*Yu et al., 2012*). Figure 3-1 illustrates the GFG platform, using one example that comprises 3 AUs with different temporal dynamics.

Reverse Correlation Experiment. On each experimental trial, the GFG randomly sampled 42 AUs and their 6 temporal parameters to synthesize random expressive face animations. The observer categorized the random facial animation according to the six classic emotion categories – ‘happy,’ ‘surprise,’ ‘fear,’ ‘disgust,’ ‘anger,’ ‘sad’ – only when the random facial movements corresponded with their mental representation of one of the emotions. Alternatively, observers selected ‘don’t know.’

Dynamic Models. To model the pattern of facial movement that drives each emotion categorization, the researchers performed a Pearson correlation between the binary activation parameter of each AU (i.e., on vs. off across trials) and the binary response variable (e.g., happy vs. not happy across trials, for happy model reconstructing), independently for each emotion. This built a 42-dimensional binary vector, per emotion and participant, specifying the contribution of each AU to a

particular emotion category. To further model the dynamic characteristics of the significant/active AUs underlying each emotion, the researchers performed, for each active AU, a linear regression between binary emotion response variables and the six temporal parameters. As a result, they computed a total of 720 dynamic facial expression models (6 expressions × male/female face × 60 participants).

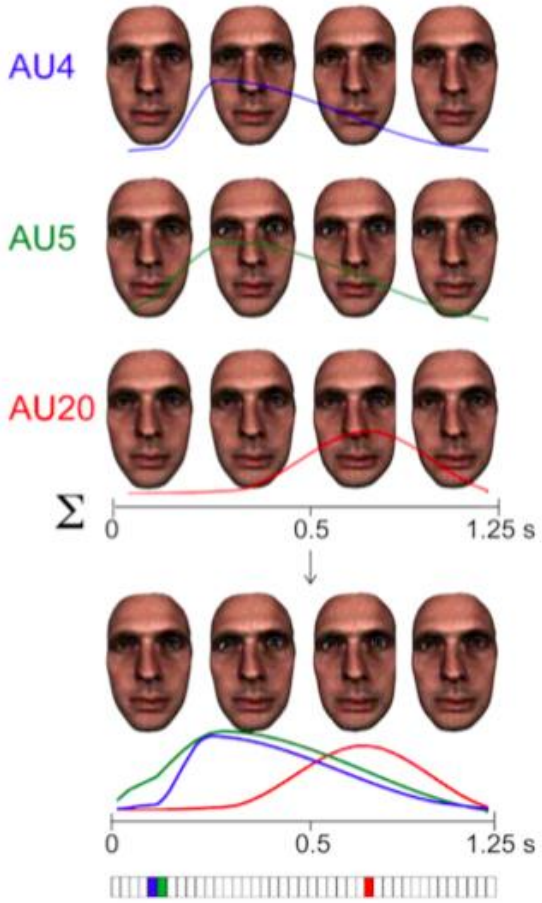

Figure 3–1 Generative Face Grammar (GFG). GFG randomly selects a subset AUs (AU4 colored in blue, AU5 colored in green and AU20 colored in red), from a total 42, and assigns values to 6 temporal parameters for each (see corresponding colored curves). These dynamic AUs are then combined and synthesized on a 3D face to produce an animation. Here, only 4 snapshots of the animation are included for the illustrative purpose. This figure is adapted from (*Jack et al., 2014*), and permission to reproduce this has been granted by the publisher Elsevier.

3.2.3 Procedure

On each trial of the validation experiment, naïve validators viewed the animation of a dynamic facial expression model and categorized it according to the six classic emotions— 'happy,' 'surprise,' 'fear,' 'disgust,' 'anger,' and 'sad.' Stimuli were presented stimuli on a black background monitor (75Hz refresh rate), in the center of the validator's visual field. A chin rest was used to ensure a constant

viewing distance of 71 cm with stimuli subtending 14.25° (vertical) and 10.08° (horizontal) of visual angle, thereby presenting faces typical of natural social interactions (*Hall, 1966; Ibrahimagić-Šeper, Čelebić, Petričević, & Selimović, 2006*). Each animation was played once for 1.25 seconds before the response options appeared and remained on the screen until the validator responded. Each validator completed 200 trials randomly sampled (with replacement) from the set of 36,000 stimuli. Each dynamic model, on average, was tested on 83 validators.

3.2.4 Analysis & Results

3.2.4.1 Categorization Accuracy

For each dynamic model, I computed the categorization performance (i.e., proportion correct to target emotion and proportion confusions to other emotions), across all tested validators. As shown in Figure 3-2, the categorization performance of these models reveals a pattern similar to that reported in experiments on static images: surprise-fear and disgust-anger elicit high confusions (see values in red in right panel), whereas happy stimuli are least confused with other emotions. Table S3-1 (see ***3.4.2***. *Supplemental Tables*) and Figure S3-1 (see ***3.4.3*** *Supplemental Figures*) showed the categorization performance of the most accurate models. This performance level together with the reported confusion pattern demonstrate that the reverse correlated dynamic models of facial expressions of emotions (which reveal the mental representation of individual observers) can in turn be used as faithful facial signals of these emotions.

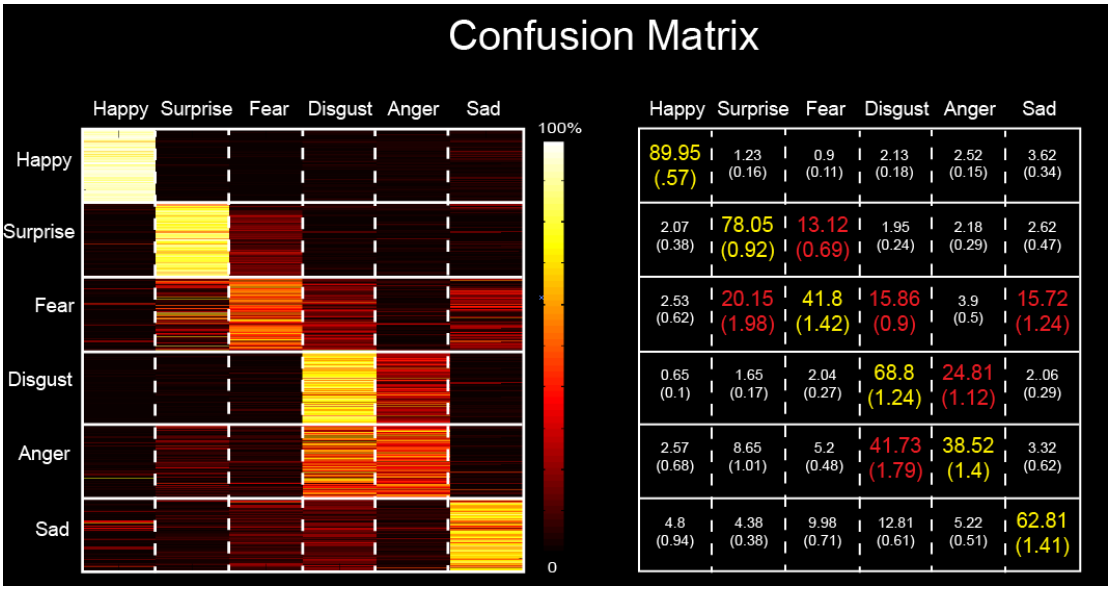

Figure 3–2 Categorization Performance of Dynamic Models. Left Panel. The color-coded confusion matrix shows the accuracy of each model used as stimulus (Y-axis) according to the six emotion categories (X-axis, categorization responses). Black-red-white color scale denotes the response proportion (black = 0%, white = 100%). Right Panel. Values in each cell show the corresponding categorization accuracy averaged across all expression models in each stimulus category (%). Yellow-colored values denote the average categorization accuracy per emotion category; red-colored values denote the high confusions (e.g., Surprise and Fear, Disgust and Anger). Standard error (SE) is between parentheses.

3.2.4.2 Analyse the Model Variants in Each Emotion

There are signal variants for each of the emotions *(Ekman & Friesen, 1978*). To estimate the main variants of these dynamic models, for each emotion I restricted the analyses to the models that were best categorized by the validators (i.e., accuracy rank >= 75 percentile across 120 models, leading to 30 models per category and in total 180 models, see Table S3-1 *in* ***3.4.2*** *Supplemental Tables*). Here, I am only interested in active AU patterns, so each model is coded as a 42-dimensional binary AU vector (i.e., AUs' on or off of each model).

To examine the structure of dynamic models in each emotion, I computed the similarity (Pearson correlation) between each pair of models using the binary AU vector, and derived for each emotion their similarity matrix. As shown in Figure 3-3, the similarity matrices of 'happy' and 'surprise' show clear, large clusters, suggesting few variations in the models of these emotions; whereas for 'fear', 'disgust', 'anger' and 'sad', their similarity matrices reveal more small clusters, suggesting more variations for these emotions.

To quantify the numbers of variants per emotion, I applied k-means clustering analysis (with k = 1 to 20) to the models, separately for each emotion. For each value of k, I computed a measure of fitness (Sum of Squared Error, SSE – the sum of the squared distance between each member of the cluster and its centroid). In each emotion category, I chose the optimal k as the minimum of the second derivative of the SSE over the values of k (the elbow criterion, see Figure S3-2 in ***3.4.3*** *Supplemental Figures*). Figure 3-3 shows the cluster assignment of each model. As predicted, models within each cluster show high similarity, on which basis I explained them as one variant. I derived the probability of each variant using the number of models they included.

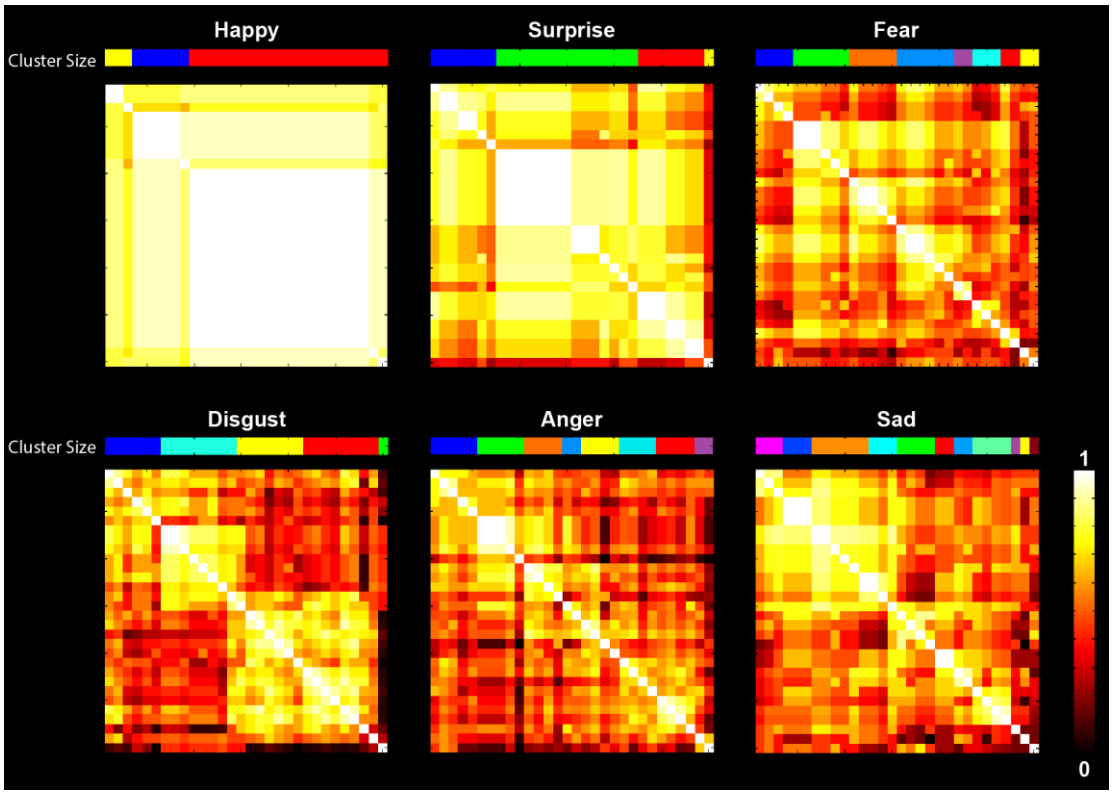

Figure 3–3 Similarity Matrix between Models in Each Emotion. Bright white indicates high similarity, whereas dark black denotes low similarity. The colored bar on the top of each similarity matrix provides the cluster assignment of each model, where each color represents a different cluster (i.e., variant), and the width of each colored bar represents the probability of each cluster (i.e., variant).

As I explained at the beginning, each dynamic model represents the mental representation of one observer. Nevertheless, the similarity across models showed in Figure 3-3 illustrates that the individual differences in the mental representations of facial expressions are not random. Rather, they form clusters with a probability of occurrence over the population. Figure 3-4 plots in pink the frequency distribution of models within each emotion variant, together with their

categorization performance as a boxplot. As shown, some AU combinations are more likely to be produced, and can lead to the typical performance reported in the literature. To illustrate the variants, I stacked the dynamic model that best fits (highest correlates) the centroid of the cluster over the bar.

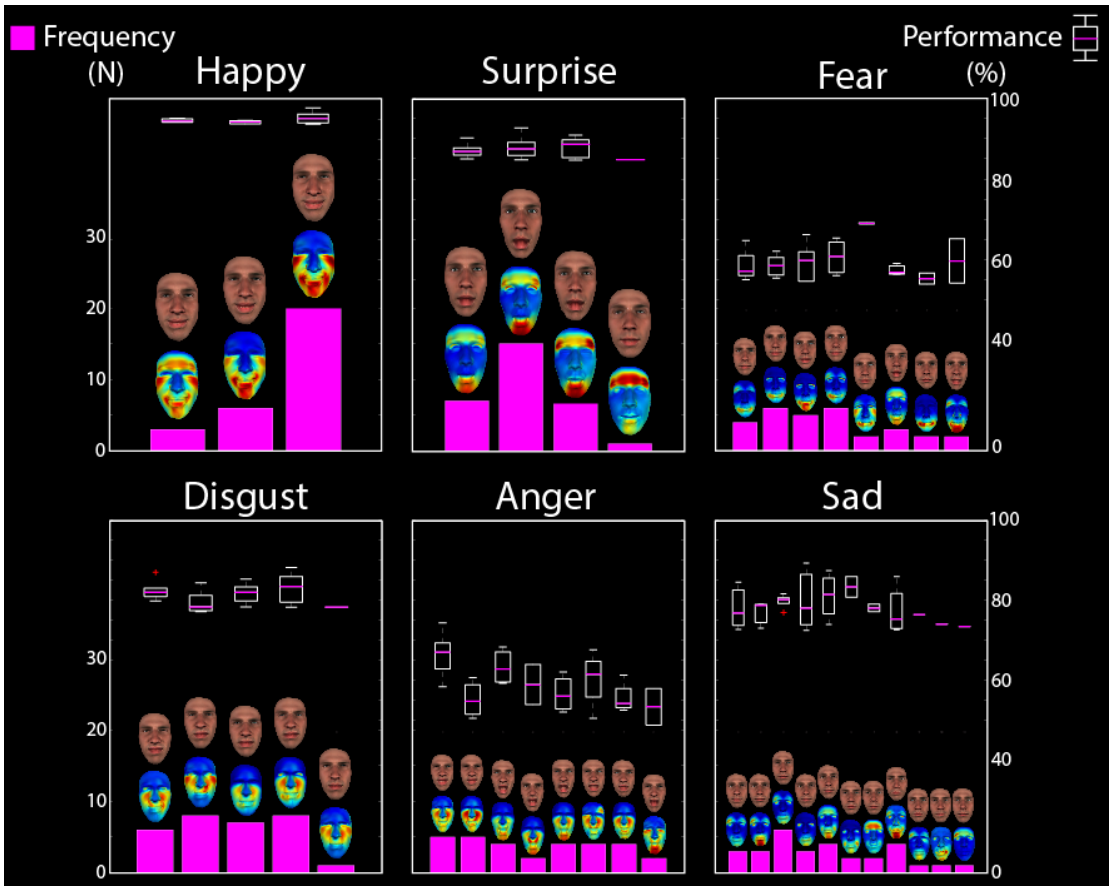

Figure 3–4 Model variants of each emotion category. In each emotion panel, pink bars plot the number of models in each variant, together with their categorization accuracy as a boxplot. The faces stacked above the frequency bars show the emotional expressions. The texture maps show the appearance of emotional expressions on an exemplary face; the color-coded heat maps illustrate the location of the active AUs, with the red colour for the highest magnitude of vertex movement.

3.2.4.3 Analyse the Model Variants and Their Probabilities as the Diagnostic Information for Facial Expressions of Emotion Recognition

Perceptual processes reflect the interplay between bottom-up extraction of information from sensory inputs and top-down inferential processes based on the mental representation that guides what information (i.e., diagnostic information) to extract from the input. If, within a culture, observers have adjusted their prior knowledge to reflect the probability of different variants in each emotion category, then their categorizations should integrate the probabilities as predictions about the likelihood of facial expression inputs. Specifically, individual observers should

know what the main variants of each emotion category are, and how likely they are as a facial expression of this category. Under this hypothesis, I can now construct a categorization model of six emotions that uses the probabilities of occurrence of emotion variants.

I built a Bayesian classifier on the 39 AFC task (one per variant category), using the probabilities of variants within each emotion as priors and setting the probability of occurrence of each emotion according to participants' response distribution across six emotions during reverse correlation (see Table S3-2 *in* ***3.4.2*** *Supplemental Tables*). In a leave-one-out method, at each of 10000 iterations, I trained a Bayesian classifier on 179 models randomly selected from the 180 models (i.e., training models) and tested its classification of the remaining model (i.e., testing model) to each of 39 response category. Each model is coded as a 42-dimensional binary AUs vector that characterizes the active AU patterns, not their temporal dynamics. Intuitively, human will perceptually ignore some AUs during recognition in reality; therefore, on each iteration I simplified the testing model by keeping the status of core active AUs (upper lid raiser and nose wrinkle *(Jack et al., 2014)*) while randomly turning off the remaining proportionally (i.e., 0%, 25%, 50%, 75% AUs off). On each iteration, I computed classification performance across all six response categories by summing the posterior probabilities across the variants of each category. For each facial expression model (row in Figure 3-5B), I averaged (N = 10000 iterations) these sums and reported them for each of 6 response categories (column in Figure 3-5B).

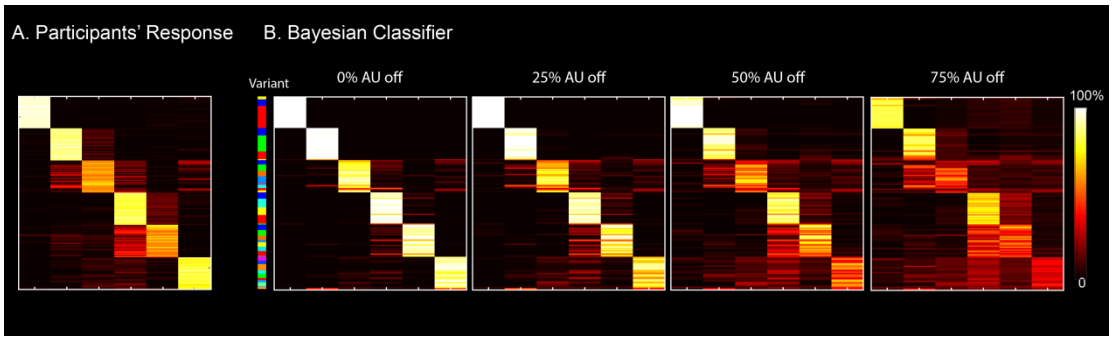

Figure 3–5 Categorization Performance of Human Validators and Bayesian Classifiers. A. Performance of Human Validators. The color-coded confusion matrix shows the accuracy of each model used as stimulus (rows) according to the six emotion categories (columns, categorization responses). B. Performance of Bayesian Classifier tested on the models with active AUs turning off proportionally, shown in the same format as A. Black-red-white color scale denotes the response proportion (black = 0%, white = 100%).

The color-coded confusion matrices in Figure 3-5 report the performance of human validators (A) and Bayesian classifier (B). As shown, the Bayesian classifier could closely replicate validators' confusion pattern for six emotions in certain case: 1) cross confusion between surprise and fear, and between disgust and anger; 2) spread confusion-prone of fear and sad; 3) and the least confusion of happy models.

To quantify how well the Bayesian classifier could mimic human's perception, I calculated the Euclidean distance between their confusion matrices, separately for each emotion and each case (i.e., 0%, 25%, 50% AUs off), with the smaller distance indicating higher similarity. Red lines in Figure 3-6 plot the classifier-human similarities when I turn off AUs under different proportion condition. To establish the statistical significance for each case, I bootstrapped a null distribution of similarities as follows. On each permutation (N = 200), I repeated above Bayesian classifier training by randomly shuffling the variant label of training models; then I calculated the classifier performance on six categories at chance-level and obtained the confusion matrix; lastly, I used the chance-level confusion matrix to computed the classifier vs. human performance similarity. Across 200 permutations, I obtained a distribution of chance-level similarity. I used the percentile 0.21 of the chance distribution as the statistical threshold (Bonferroni corrected, $p < 0.05$, 1-tailed). Black dash lines in figure 3-6 show the statistical threshold.

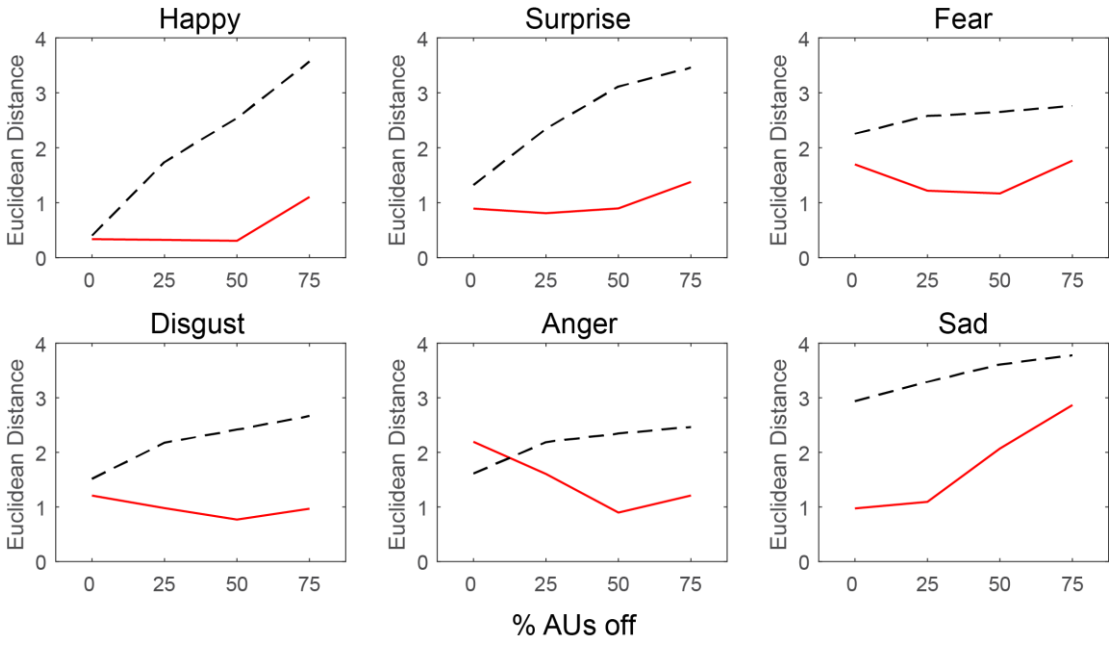

Figure 3–6 Performance Similarity between Human Vlidators and Bayesian Classifier. In each emotion panel, the red line plots the similarity (i.e., calculated by the Euclidean distance) between the confusion matrices of human observer and Bayesian classifier in each case of proportional AUs off.

Figure 3-6 shows that the Euclidean distances between the performance of human and the Bayesian classifier are closer than chance in most case, except the anger models tested by keeping all active AUs. These results demonstrate that the main signal variants and their probability I have derived work well to depict the expectation of individuals in Western population. Also, the Bayesian classifier mimics human performance best when it is tested on models with some active AUs turned off (i.e., 50% off for Happy, Fear, Disgust, Anger, and 25% off for Surprise), indicating that human does not take all active AUs into account during their recognition but bias towards fewer AUs.

3.3 Discussion

In this study, I tested 720 dynamic models of facial expression of six emotions on a group of WC participants and demonstrated the validity of a subset models. Clustering analysis on the validated models revealed main signal variants for each emotion, and showed some variants are more likely to be produced than others. The Bayesian classifier was then trained using these variants and their probabilities, which showed a level of categorization performance mimics human observers closely. Together, my results demonstrate that 1) these reverse

correlated dynamic models are sufficiently precise to characterize the diagnostic facial movements for facial emotions recognition in the Western population; 2) they can be structured as some main signal variants for each emotion category; 3) the probabilities of these variants should have been learned by Western people and construct their mental representation of facial expressions of emotion.

It should be noted that, though the reverse-correlated models have high resolution on their temporal dynamics (i.e., modelled by 6 temporal parameters), my analysis tapped in only their spatial pattern (i.e., AUs binary on-off pattern). Future work need to clarify the psychological status of AUs' dynamics. At this stage, it would like to step away from current results and emphasize the potentials of using GFG platform to study the temporal dynamics of facial movement.

Facial expressions are by nature dynamic, thus understanding the mental representation of facial expressions requires the incorporation of their dynamic properties. With GFG platform, we can rigorously control such temporal dynamics at very high temporal/spatial resolution by precisely specifying the active AUs and their six temporal parameters. Such manipulation goes beyond the current state of the art, which creates dynamic face by either proportionally amplifying the intensity of AUs at each frame using the morphing techniques (arbitrary temporal properties, e.g. *Kamachi et al., 2001*and *Krumhuber & Kappas, 2005*), or recording facial expressions generated by instructed actors (no manipulable temporal properties, e.g. *Kanade, Cohn, & Tian, 2000; van der Schalk, Hawk, Fischer, & Doosje, 2011; L. J. Yin, Chen, Sun, Worm, & Reale, 2008*). Together with the reverse correlation approach, the GFG platform provides a powerful tool to decipher the psychological relevance of different AUs and their dynamic properties. For example, by analysing reverse correlated dynamic AUs patterns, researchers have demonstrated that the biologically rooted facial signals (e.g., Upper Lid Raiser and Nose Wrinkler which modulate sensory exposure, see *Rozin & Fallon, 1987; Rozin, Lowery, & Ebert, 1994; Susskind et al., 2008*) are transmitted earlier to express elementary categories (e.g., approach/avoidance); whereas, more complex signals are transmitted later for the discrimination of six emotions (*Jack et al., 2014*). The same approach can be applied to a broad spectrum of facial expression, ranging from very fundamental signals (i.e., physical pain and pleasure) to very socially-interactive signals (i.e., mental states), to fully understand the functional role of the facial signal dynamics.

3.4 Supplemental Materials

3.4.1 Screening Questionnaire

I selected participants who answered ‘NO’ to ALL questions.

Have you ever:

i) lived in a non-Western country before (e.g., on a gap year, summer work, move due to parental employment)?*

ii) visited a non-Western country (e.g., vacation)?

iii) dated or had very close friendship with a non-Western person?

iiii) been involved with any non-Western cultural societies/groups?

*by Western country/person/group, we are referring to Europe (East and West), North American, Australia, and New Zealand.

3.4.2 Supplemental Tables

Table S 3-1 Confusion matrix of models with categorization accuracy ranking over 75%. Each cell shows the averaged response proportion (%) of models in each emotion category (row) to six emotion labels (column). SE is in *Italic*.

		Happy	Surprise	Fear	Disgust	Anger	Sad
	Happy	**94.87**	0.51	0.51	0.86	1.56	1.69
		-0.2	*-0.12*	*-0.13*	*-0.14*	*-0.17*	*-0.25*
	Surprise	1.07	**87.54**	8.76	0.55	0.87	1.21
		-0.22	*-0.4*	*-0.6*	*-0.19*	*-0.22*	*-0.34*
%	Fear	0.75	15.13	**59.45**	13.71	1.41	9.56
		-0.19	*-2.05*	*-0.82*	*-1.37*	*-0.26*	*-1.64*
	Disgust	0.36	0.92	1.68	**81.45**	13.27	2.32
		-0.11	*-0.16*	*-0.27*	*-0.57*	*-0.53*	*-0.36*
	Anger	1.23	5.49	3.99	29.38	**59.25**	0.66
		-0.4	*-0.75*	*-0.53*	*-1.77*	*-1.16*	*-0.18*
	Sad	0.55	2.78	6.01	8.9	3	**78.76**
		-0.16	*-0.42*	*-0.69*	*-0.74*	*-0.42*	*-0.85*

Table S 3-2 Number of responses in each of 6 emotion categories.

	Happy	Surprise	Fear	Disgust	Anger	Sad
Mean	355.45	338.63	244.78	451.7	358.93	279.42
SD	150.72	112.41	114.61	152.46	158.06	131.28

SD = standard deviation.

3.4.3 Supplemental Figures

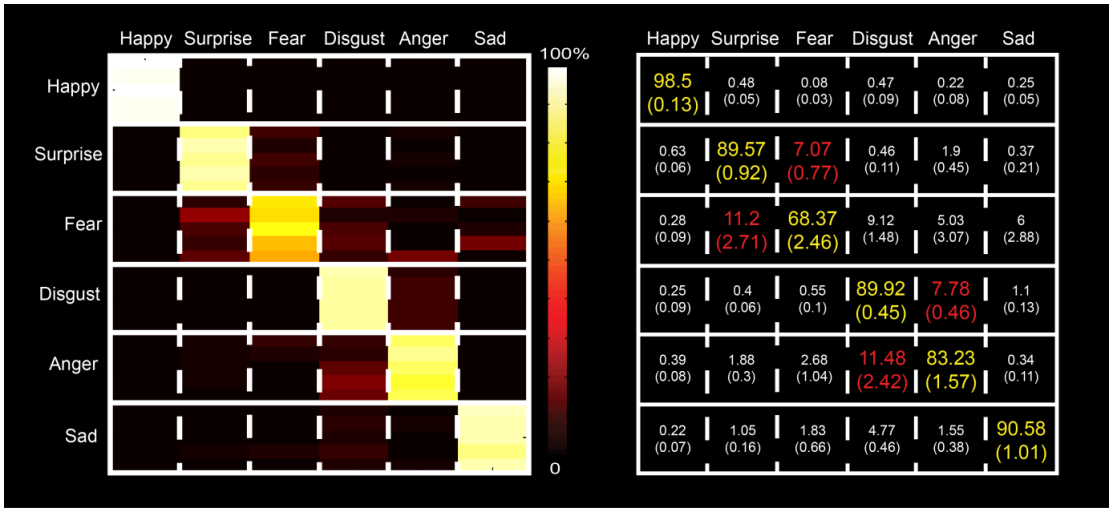

Figure S 3-1 Categorization Performance of the Best Dynamic Models. Five models with the best performance in each emotion category are selected (N = 30 in total). Their performances are illustrated in the same format as Figure 3-1.

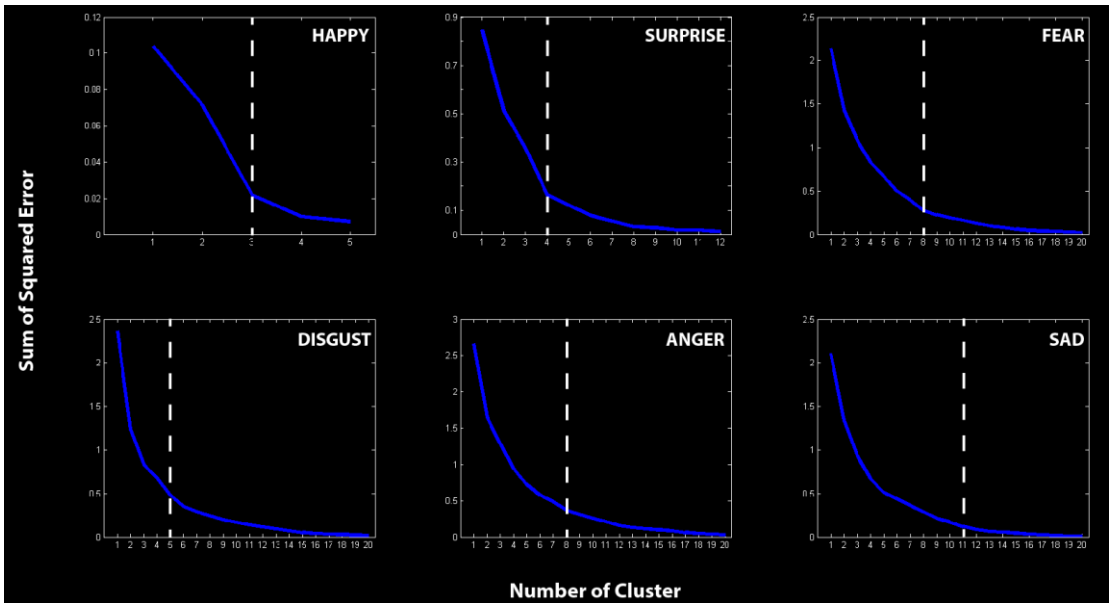

Figure S 3-2 Number of Clusters. In each emotion category, the blue line plots the sum of squared distance between each member of the cluster and its centroid (Y-axis), calculated across a range of cluster numbers (X-axis, i.e., k=1-20, inclusively). The white dash line shows an optimal number of k, where adding more clusters cannot explain considerable variance more.

4 Study 3: Dynamic Construction of Diagnostic Information in the Brain for Perceptual Decision Behavior

4.1 Introduction

Over the past decade, extensive studies of the brain regions that support face, object and scene recognition suggest that these regions have a hierarchically organized architecture that spans the occipital and temporal lobes *(Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; K. Friston, 2008; Grill-Spector & Weiner, 2014; Guclu & van Gerven, 2015; Kravitz et al., 2013; Riesenhuber & Poggio, 1999; Schwiedrzik & Freiwald, 2017; Sigala & Logothetis, 2002; Van Essen et al., 1992)*, where visual categorizations unfold over the first 250 milliseconds of processing *(Cichy, Pantazis, & Oliva, 2014; Ince et al., 2016; Liu, Harris, & Kanwisher, 2002; Schyns, Petro, & Smith, 2007; VanRullen & Thorpe, 2001)*. This same architecture is flexibly involved in multiple tasks that require task-specific diagnostic representations—e.g., categorize the same face as “happy,” “Mary,” the same object as “a car” or “a Porsche,” and the same scene as “city” or “New York.” While we partly understand where and when these categorizations happen in the occipito-ventral pathway, the next challenge is to unravel *how*. That is, how does high-dimensional input collapse in the occipito-ventral pathway to become low dimensional representations (i.e., the diagnostic information) that guide behavior? To address this, in this study, I investigated what information the brain processes in a scene categorization task and visualized the dynamic representation of this information in brain activity.

To do so, I used the Contentful Brain and Behavior Imaging (CBBI), an information theoretic framework, to tease apart stimulus information that supports behavior (i.e., diagnostic) from that which does not (i.e., nondignostic). I then tracked the dynamic representations of both in magneto-encephalographic (MEG) activity. Using CBBI, I demonstrated that a rapid (~170 ms) reduction of behaviorally irrelevant information occurs in the occipital cortex and that representations of the information that supports distinct behaviors are progressed ventrally and constructed in the right fusiform gyrus (rFG).

4.2 Experiment

4.2.1 Participants

Five participants with normal (or corrected to normal) vision participated in the experiment. We obtained informed consent from all participants and ethical approval from the University of Glasgow Faculty of Information and Mathematical Sciences Ethics Committee.

4.2.2 Stimuli

I cropped a copy of Dali's *Slave Market with the Disappearing Bust of Voltaire* (see Figure 4-1 A-a, Stimulus) because it contains a complex, ambiguous scene that observers perceive as either "the nuns" or "Voltaire." The cropped image size was 256 x 256 pixels, presented at 5.72° × 5.72° of visual angle on a projector screen. I used the Bubbles technique (*Gosselin & Schyns, 2001*) to break down the stimulus information into random samples for each experimental trial (see Figure 4-1 A-a, Stimulus Sampling). I now explain the sampling procedure in details.

I first decomposed the image into six independent Spatial Frequency (SF) bands of one octave each, with cut-offs at 128 (22.4), 64 (11.2), 32(5.6), 16 (2.8), 8 (1.4), 4 (0.7) cycles per image (c/deg of visual angle), respectively. For each of the first five SF bands, a bubble mask was generated from a number of randomly located Gaussian apertures (the bubbles), with standard deviations of 0.13, 0.27, 0.54, 1.08, and 2.15 deg of visual angle, respectively. I sampled the image content of each SF band by multiplying the bubble masks and underlying greyscale pixels at that SF band, summed the resulting pixel values across SF bands, and added the constant 6^{th} SF band to generate the actual stimulus image. The total number of 60 Gaussian apertures on each trial remained constant throughout the task, ensuring that equivalent amounts of visual information were presented for each trial, at a level found previously to maintain "don't know" responses at 25% of the total response number (*Schyns, Jentzsch, Johnson, Schweinberger, & Gosselin, 2003*). Since the 6^{th} underlying SF image was constant across trials, I performed all analyses on the 5 bubble masks controlling visibility, but reported only the first three because they represented most of the information required for perceptual decisions. For analysis, I down-sampled (bilinear interpolation) the bubble masks to a resolution of 64 × 64 pixels to speed up computation.

4.2.3 Procedure

Each trial started with a fixation cross displayed for 500 ms at the centre of the screen, immediately followed by a stimulus generated as explained above that remained until response. We instructed observers to maintain fixation during each trial, and to respond by pressing one of three keys ascribed to each response choice—i.e. “the nuns”, “Voltaire”, or “don’t know” (see Figure 4-1 A-b, Perceptual Decision). Stimuli were presented in runs of 150 trials, with randomized inter-trial intervals of 1.5–3.5s (mean 2s). Observers performed 4–5 runs in a single day session with short breaks between runs. Observers completed the experiment over 4–5 days.

4.2.4 MEG Data Acquisition

I measured the observers’ MEG activity with a 248-magnetometer whole-head system (MAGNES 3600 WH, 4-D Neuroimaging) at a 508 Hz sampling rate. I performed analysis with the FieldTrip toolbox (*Oostenveld, Fries, Maris, & Schoffelen, 2011*) and in-house MATLAB code, according to recommended guidelines *(Gross et al., 2013)*. For each participant, I discarded runs based on outlying gradiometer positions in head-space coordinates. That is, I computed the Mahalanobis distance of each sensor position on each run from the distribution of positions of that sensor across all other runs. Runs with high average Malahanobis distance were considered outliers and removed. The distances were then computed again and the selection procedure was repeated until there were no outlier runs (Mahalonobis distances > 20). I high-passed filtered data at 1 Hz (4^{th} order two-pass Butterworth IIR filter), filtered for line noise (notch filter in frequency space) and de-noised via a PCA projection of the reference channels. I identified noisy channels, jumps and other signal artefacts using a combination of automated techniques and visual inspection. I then epoched the resulting data set (mean trials per observer 3396, range 2885–4154, see Table S4-1) into trial windows (−0.8s to 0.8s around stimulus onset) and decomposed using ICA, separately for each observer. I identified and projected out of the data the ICA sources corresponding to artefacts (eye movements, heartbeat; 3 to 4 components per observer).

I then low-pass filtered the data to 40Hz (3^{rd} order Butterworth IIR filter), specified our interest time period 0-400ms post stimulus, and performed the Linearly Constrained Minimum Variance Beamforming analysis (*VanVeen,*

vanDrongelen, Yuchtman, & Suzuki, 1997) to obtain the source representation of the MEG data on a 6mm uniform grid warped to standardized MNI coordinate space (12,773 sources, henceforth I call them MEG voxels). I low-pass filtered the resulting single trial voxel time courses with a cut-off of 25Hz (3^{rd} order Butterworth IIR filter, two-pass). In the following analysis, based on the obtained single trial voxel activity time courses (12,773 MEG voxels, every 2ms between 0 - 400ms post stimulus), I analyzed the dynamic representation of features in the brain for perceptual decisions.

The following sections detail each step of the information processing pipeline. Figure S4-14 provides a schematic graphic overview of the pipeline.

4.2.5 Analysis & Results

4.2.5.1 Contentful Brain and Behavior Imaging (CBBI)

Our colleague developed the CBBI based on the statistical framework of information theory for neural imaging data analysis. This framework can measure the co-representations between three essential components, i.e. stimulus information, brain activity, behavioral response, and therefore enables us to understand the information processing from brain activity. In this section, I will briefly introduce this information theoretical framework, which estimate is mainly based on the Mutual Information (MI) and Redundancy (Red).

MI measures the statistical dependence between two variables (*Cover & Thomas, 1991; Shannon, 1948*), by calculating entropy differences. As entropy quantifies the uncertainty of variables, MI actually measures the reduced uncertainty (or increased certainty) about variable X based on the knowledge of variable Y, which is in other words the common variations between two variables. We can define the MI in three mathematically equivalent ways:

$$MI\langle R; F\rangle = H(F) - H(F|R) \quad (1)$$

$$= H(R) - H(R|F) \quad (2)$$

$$= H(R) + H(F) - H(R, F) \quad (3)$$

To illustrate in a psychology context, $MI\langle R; F\rangle$ is the MI between the response distribution R and the stimulus feature distribution F. In the formula, $H(F)$ denotes the entropy of feature distribution F, and $H(R)$ is the entropy of response

distribution *R*. *H*(*F*|*R*) represents the conditional entropy, i.e. the entropy of feature distribution *F* on the presence of response value r of *R*. *H*(*R*,*F*) is the entropy of the joint distribution of *F* and *R*.

In (1), MI quantifies the average reduction in uncertainty about which feature was presented when we observed a response of *R*. Symmetrically, MI defined in (2) quantifies the average reduction in uncertainty about the response when we know which stimulus feature is represented. In (3), MI quantifies the difference between the entropy of a model in which features representations and responses are hypothesized as statistically independent and the entropy of their true joint distribution. This should be “the most useful interpretation of MI from neuroimaging perspective: a statistical test for independence.” (*Ince et al., 2017*)

MI has several useful properties. First, its calculation requires no assumptions on the distribution of variables, and therefore can quantify non-parametrically the relationship between variables (i.e. linear vs. nonlinear). Second, MI is additive for independent variables: $MI <R_1, R_2; F> = MI <R_1; F> + MI<R_2; F>$, which is derived from the logarithm of entropy calculation (see detailed explanation in (*Ince et al., 2017*)). The addictive property is crucial because it enables the quantification of triple interaction effects, termed the redundancy (Panzeri, Magri, & Logothetis, 2008; Schneidman, Bialek, & Berry, 2003; Timme, Alford, Flecker, & Beggs, 2014), by calculating the MI which is shared between $MI<R_1; F>$ and $MI <R_2; F>$:

$$Red <R_1, R_2, F> = MI <R_1; F> + MI <R_2; F> – MI <R_1, R_2; F> \quad (4)$$

In (4), Red quantifies the difference between the MI of a model in which two responses of feature representations are hypothesized as statistically independent and the MI of considering both Reponses together. This measures the statistic independence of three variables, or alternatively the independence of two co-represented information.

In my thesis, I use “< ; >” to denote the relationship between variables measured by the information theoretic statistics introduced above - MI and Red.

4.2.5.2 Diagnostic Features of Behavior

To compute the diagnostic features of perceptual decisions, I quantified the statistical dependence between the pair <Information Samples; Perceptual

Decision> using Mutual Information (MI, *Ince et al., 2017*). I used MI because it non-parametrically quantifies the common variations between information and decisions to reveal the features that support the decision. I schematized the relationship between the two variables of information sample and decision as a Venn diagram in figure 4-1A, where they intersect was designated the "diagnostic features" that support each observer's decisions (*Schyns, 1998; Tversky, 1977*). Using MI, I computed diagnostic features separately for the behavioral contrasts <Information Samples; "the nuns," vs. "don't know">, excluding "Voltaire" trials, and <Information Samples; "Voltaire," vs. "don't know">, excluding "nuns" trials. I now explain the calculations in details. I precede the MI calculations in 3 steps:

Step 1: Binarize the Pixel Visibility. On each trial, 5 real-valued Gaussian bubble masks multiply the visual information represented in 5 SF bands (see Figure 4-1 A-a, Stimulus Sampling, for an illustration). Thus, on a given trial, a real value represents the visibility of that pixel under a Gaussian bubble, with 1 indicating full visibility and 0 indicating no visibility. For each pixel of the bubble mask, I converted its random distribution of real values across trials into 2 bins—values below 0.2 were ascribed to the "no to low visibility" bin and values above 0.2 to the "low to full visibility" bin.

Step 2: MI <Pixel Visibility; Perceptual Decision>. I used MI to quantify the statistical dependence between the binarized pixel visibility values and the corresponding observer responses, grouping "the nuns" vs. "don't know" responses together in one computation (i.e. <Pixel Visibility; "the nuns," "don't know">) and the "Voltaire" vs. "don't know" responses in the other (i.e. <Pixel Visibility; "Voltaire," "don't know">).

Step 3: Diagnostic Pixels for Each Perception. Computations in step 2 resulted in two MI perceptual decision pixel images per participant (see Figure S4-1A in ***4.4.3*** *Supplemental Figures* for the thresholded classification images for each participant). I used the method of maximum statistics (*Nichols & Holmes, 2002*) to determine the statistical significance of MI pixels and correct for multiple comparisons. Specifically, for each of 10,000 permutations, I randomly shuffled the participant's choice responses across trials, repeated the computation of MI for each pixel as explained and extracted the maximum MI across all pixels over the 5 SF bands. I used the 99^{th} percentile of the distribution of maxes across 10,000

permutations to determine the above-chance significance of each MI pixel (FWER $p < 0.01$, one-tailed).

Across observers, I reported the diagnostic pixels with significant MI in the first 3 SF bands that illustrate the consistency of the main diagnostic features underlying perceptual decision behaviors (see Figure 4-1A-c, Diagnostic Features of Behavior). As shown in Figure 4-1A-c, all participants used the left and right nun's faces at higher spatial frequencies (HSF) to respond "the nuns", whereas they used the global face of Voltaire at lower spatial frequencies (LSF) to respond "Voltaire" (see Figure S4-1A in ***4.4.3*** *Supplemental Figures* for each participant's features). Since diagnostic features influence behavior, the participant's brain must represent at a minimum these features between stimulus onset and observer decision. Next, I will show that the brain does indeed represent all diagnostic features over time, as well as other features.

4.2.5.3 Representation of Features in the Brain

To show where and when each participant's MEG activity represents stimulus features, I used MI to evaluate the single-trial relationship <Information Samples; MEG Voxel Activity>. Here, I used the Gaussian-Copula Mutual Information estimator (*Ince et al., 2017*) for continuous values.

In each participant, I measured single-trial MEG activity with the bivariate of amplitude and instantaneous MEG gradient on 12,773 voxels, every 2 ms between 0 and 400 ms post-stimulus. A high-dimensional 12,773 x 200 voxel-by-time matrix therefore structures the MEG data. For each participant, I aimed to quantify the features of the stimulus that each cell of this matrix represents, if any. I proceeded in three steps.

Step 1: Computation of the Relationship <Information Samples; MEG Activity>. I aim to identify, in each participant, the features represented in each cell of the full voxel-by-time matrix of MEG activity. However, it is computationally impractical to directly compute the features from the single-trial relationship <Information Samples; MEG Voxel Activity>, due to the enormous dimensionality of the space—64 x 64 x 5 SF bands pixels x 12,773 voxels × 200 time points. Instead, I used the method reported in *(Ince et al., 2015)*, which computes the relationship over the more computationally tractable matrix of 60 Independent

Component Analysis (ICA) sources representing MEG activity over 75 time points that span 0 to 600 ms post stimulus every 8 ms.

Step 2: Computation of Brain Features. For each participant, the reduced matrix computed above (i.e., 60 ICA sources x 75 time points) comprised MI images in each cell, for a total of 4,500 MEG-pixel information images across 5 SF bands. I vectorized each (64 x 64 x 5 = 200,480) MEG MI image as a 20,480-dimensional vector. I applied Non-negative Matrix Factorization (NMF, *D. D. Lee & Seung, 1999*) to the set of 4,500 vectorized MEG MI images to characterize the main NMF features of the stimulus that modulate MEG source activity, resulting in 21–25 components per observer. I thresholded these NMF features by setting to zero the pixels with low MI values (< 15% of the maximum pixel value across SFs). I then normalized the NMF features (L2-norm). Henceforth, we call "brain features" the normalized NMF features of each participant that modulate the MEG activity of their brain.

Step 3: Computation of the Relationship < Brain Feature; MEG Voxel Activity> in the Full Voxel-by-Time MEG Activity Matrix. I used the brain features computed above from the reduced matrix of ICA MEG sources to quantify their representation into each cell of the full voxel-by-time matrix. To this aim, first I computed the visibility of each brain feature into the information samples (i.e., bubble mask) presented as stimulus on each trial. That is, I spatially filtered (i.e., dot product) the bubble mask for that trial with the brain feature computed above, thereby producing a scalar value indicating the visibility of this feature on this trial. I call these real values "brain feature coefficients." Next, for each brain feature, and for each cell of the full voxel-by-time MEG activity matrix, with MI I quantified the relationship <Brain Feature Coefficient; MEG Voxel Activity>. This produced for each participant, a 3D feature-by-voxel-by-time MI matrix. I determined the statistical significance for each cell using a permutation approach and the method of maximum statistic to address multiple comparisons (*Nichols & Holmes, 2002*). Specifically, for each of 200 permutations, I randomly shuffled the brain feature coefficients values across trials and recalculated the MI of the single trial relationship <Randomized Brain Feature Coefficients; MEG Voxel Activity>. I then computed the maximum of the resulting 3D MI matrix for each permutation and used the 95^{th} percentile of this maximum value across permutations as the statistical threshold (i.e., FWER $p < 0.05$, one-tailed). In the remaining analyses, I

used the thresholded 3D feature-by-voxel-by-time MI matrix of each participant which I called "**Representation Matrix**." The 3D **Representation Matrices** for each participant are unique to my analysis: they reveal the stimulus features that the brain dynamically represents, separating out the features that are relevant for the perceptual task.

Figure 4-1B shows the common brain features represented cross observers. Comparing Figure 4-1B with Figure 4-1A reveals that some brain features correspond to the same visual information as the features that are diagnostic of behavior (i.e., the red and blue nun's faces at HSFs and the green face of Voltaire at LSFs), whereas others do not (e.g. the brown features flanking Voltaire's face).

4.2.5.4 Diagnostic and Nondiagnostic Brain Features.

Now I divided the brain's features into diagnostic or nondiagnostic for the task. The Venn diagram of Figure 4-1B illustrates such division: the addition of brain measures produces a white area of intersection that represents the diagnostic features that influence both behavioral and brain measures; the magenta intersection designates the nondiagnostic features that influence brain measures but not behavior.

For each participant, I determined the diagnostic vs. nondiagnostic status of their brain features as follows. Using only the trials associated with "the nuns" vs. "don't know" behavioral responses, I computed the single-trial MI relationship <Brain Feature Coefficient; "the nuns," "don't know"> to derive the brain features diagnostic for "the nuns" perception. Likewise, I computed independently the single-trial MI relationship <Brain Feature Coefficient; "Voltaire," "don't know"> to derive the brain features diagnostic for "Voltiare" perception, using only the trails including "Votaire" and "don't know" reponse. In both cases, a strong relationship (i.e., MI above 75^{th} percentile of the distribution of MI across all brain features) would classify this brain feature as diagnostic—i.e., of "the nuns" or of "Voltaire." Finally, to decide the brain features that are irrelevant to behavior, I use all trials and computed the single-trial MI relationship <Brain Feature Coefficient; "the nuns," "Voltaire," "don't know">. A weak relationship (MI below 25^{th} percentile of the MI distribution) would classify this brain feature as nondiagnostic of perceptual decisions (see Figure S4-1B in ***4.4.3*** *Supplemental Figures* for the perception-specific brain features and nondiagnostic features of each participant).

Figure 4-1C illustrates the expected topological representation of brain features during the first 20 ms of representation (see Figure S4-2 in ***4.4.3*** *Supplemental Figures* for each participant's topological representation). Color codes reveal that the participants' brains contral-laterally represented the diagnostic eyes of Voltaire (see the red and blue voxels) and the brown nondiagnostic features flanking the centre of the stimulus, in relation to the bilaterally represented LSF Voltaire face (see green voxels).

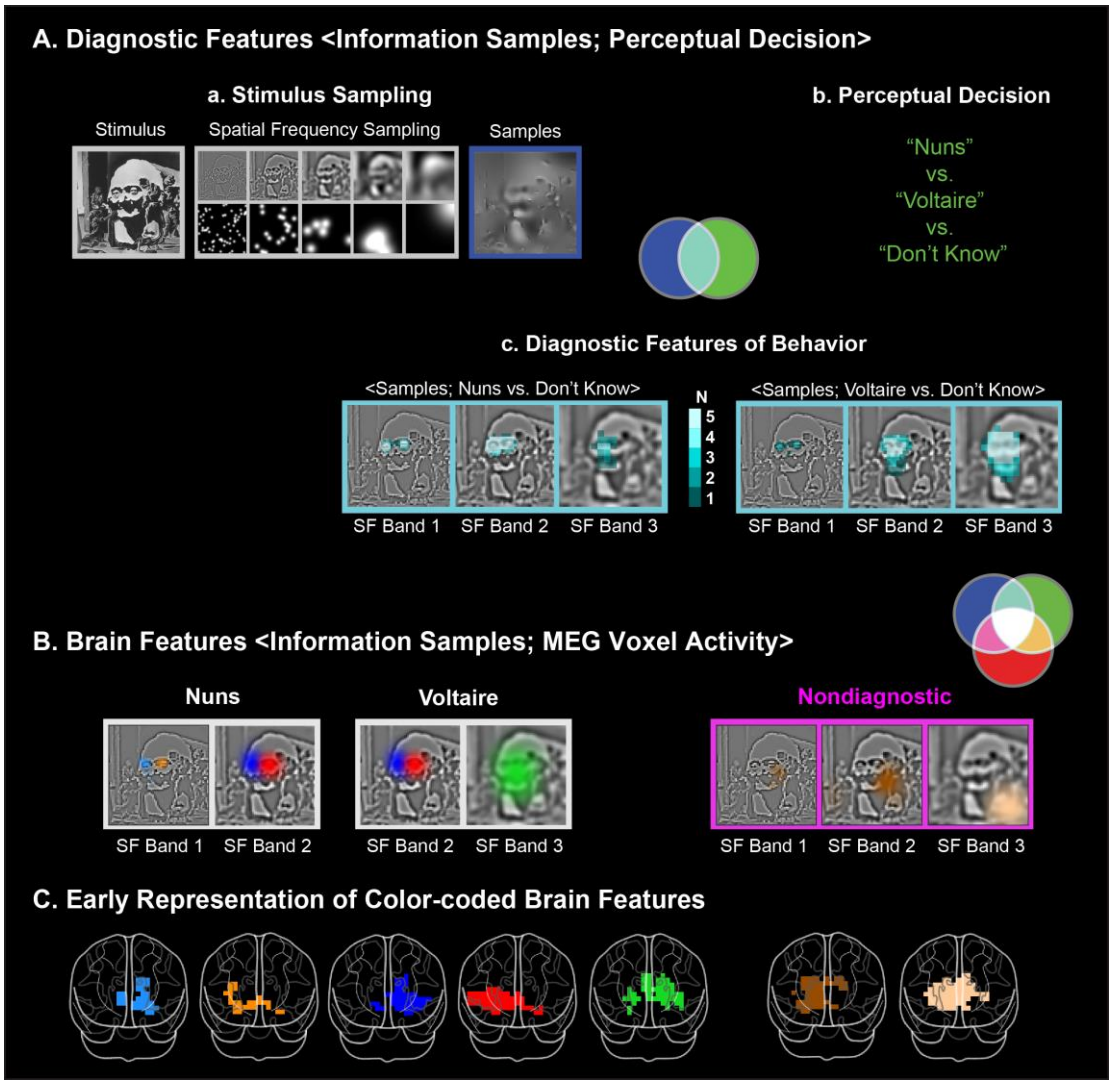

Figure 4–1 Diagnostic and Brain Features. A. Diagnostic Features. (a) The original stimulus (left), which was decomposed into 6 spatial frequency (SF) bands (middle, band 6 is not shown) of one octave each for each trial, starting at 128 cycles per image. Samples were added across bands to generate one experimental stimulus (dark blue frame, right). (b) Perceptual decisions recorded by observers, as: "the nuns," "Voltaire," or "don't know". The cyan intersection in the Venn diagram illustrates the relationship between information samples (blue) and perceptual decisions (green): the diagnostic features of behavior. (c) Diagnostic feature of behavior. The cyan-framed images show significant pixels (Family-wise error rate (FWER), $p < 0.01$, one-tailed) in the first three SF bands that reveal features diagnostic for observers responding "the nuns" (the two small

faces in SF band 1) and “Voltaire” (the broad face in SF band 3). Color saturation indicates N, number of observers. B. Brain Features. White frames highlight “the nuns” and “Voltaire” diagnostic and color-coded brain features represented by a majority of observers (i.e. N>3). The magenta frames highlight color-coded non-diagnostic brain features represented by a majority of observers (i.e. N>=3). The magenta intersection in the Venn diagram represents the relationship between information samples (blue) and MEG voxel activity (red) whereas the white intersection represents the relationship between all three variables, including behavior. (C) Early Representation of Brain Features. Common, color-coded brain regions, show the early (during the initial 20 ms of representation) topological representation of each correspondingly colored brain feature (FWER, $p < 0.05$, one-tailed). Each observer contributed at least one significant voxel for each color-coded feature.

4.2.5.5 Divergence of Nondiagnostic and Diagnostic Brain Features in the Occipito-Ventral Pathway

To examine the representation divergence of diagnostic vs. nondiagsotic brain features for each observer, I used their un-thresholded full 3D **Representation Matrix**. For each of the 5,869 cortical voxels, I extracted the max MI across all diagnostic (vs. nondiagnostic) features and all time points in 10 ms time windows between 0 and 400 ms post stimulus. This resulted in one 2D matrix (un-thresholded MI of 5869 voxels in 40 time windows) of diagnostic feature representation and another of nondiagnostric feature representation. Using this 2D matrix, in each time window, I computed the similarity between diagnostic and nondiagnostic representations with the de-meaned dot-product between the two 5,869 dimensional vectors. To establish statistical significance, I bootstrapped a null distribution as follows. On each iteration (N = 1000), I randomly shuffled the values across the dimensions of the two 5,869 dimensional vectors and calculated their de-meaned dot product. I used the percentile 0.625 and 99.9375 of the chance distribution as the upper and lower boundaries for the chance-level similarity (Bonferroni corrected, $p < 0.05$, 2-tailed). I performed the same analysis at the group level, by pooling all participants’ data together to form a larger 2D matrix (29345 voxels by 40 time windows). I found that diagnostic and nondiagnostic brain features diverge around 170 ms post-stimulus (see Figure S4-3 in ***4.4.3*** *Supplemental Figures*).

On this basis, I defined an earlier ([50-170 ms] post stimulus) and a later time window ([170-400 ms] post stimulus), that flank the N/M170; the Event Related Potential ~170 ms post-stimulus commonly associated with visual categorizations (*Bentin, Allison, Puce, Perez, & McCarthy, 1996; Cichy et al., 2014*). I summarized the representation of brain features in each window as such:

a voxel would represent diagnostic (vs. nondiagnostic) brain features if it has significant MI (FWER $p < 0.05$, one-tailed) for at least one diagnostic (vs. nondiagnostic) brain feature in this time window. For each voxel, then I counted the number of participants satisfying these criteria and reported the distributions for diagnostic (white schematic brains in Figure 4-2) and nondiagnostic (magenta schematic brains in Figure 4-2) brain features in each time window. I reported only the 5,869 cortical voxels in the figures.

The color-coded brains in Figure 4-2 summarize the evolving representations of the diagnostic and nondiagnostic features across two post-stimulus time windows. A comparison of the nondiagnostic and diagnostic brain features across the earlier and later time windows reveals a consistent pattern. Over the first 170 ms of processing, representation of diagnostic and nondiagnostic brain features similarly involve occipital cortex (Bonferroni corrected $p < 0.05$, two-tailed). They diverge afterward and only representations of diagnostic brain features are sustained in all occipito-ventral regions.

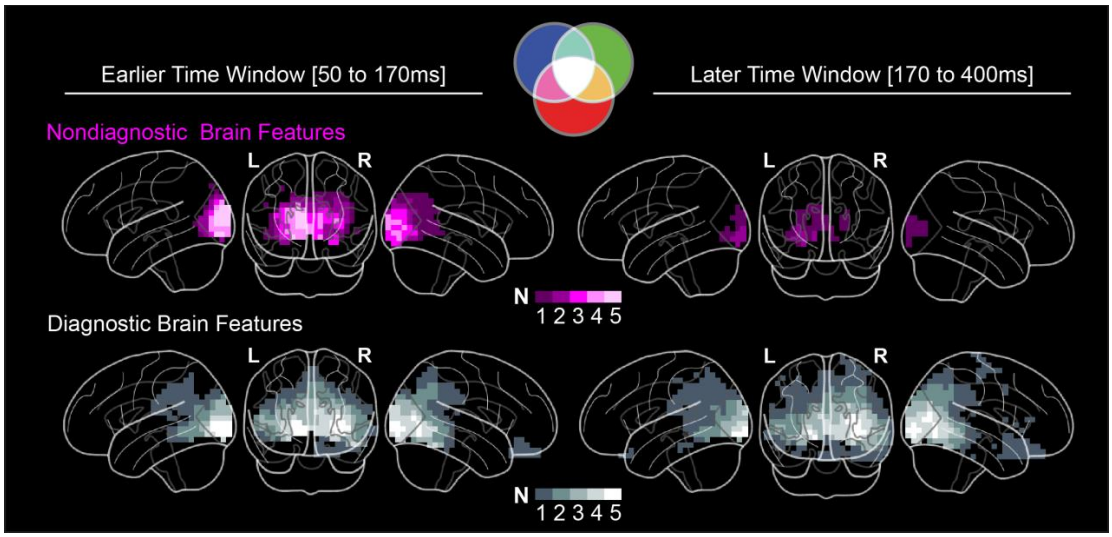

Figure 4–2 Nondiagnostic Feature Reduction and Diagnostic Feature Progression. Magenta color-coded brains show voxels that represent at least one significant (FWER $p < 0.05$, one-tailed) nondiagnostic brain feature (represented with a magenta color in the Venn diagram) in earlier [50-170 ms] and later [170-400 ms] time windows, post stimulus. White color-coded brains show voxels that represent at least one significant (FWER $p < 0.05$, one-tailed) diagnostic brain feature (represented with a white color in the Venn diagram) in earlier [50-170 ms] and later [170-400 ms] time windows, post stimulus. Voxel brightness denotes the number (N) of participants for whom these criteria held true. For all observers, nondiagnostic features were consistently reduced over time in the occipital cortex while diagnostic features were sustained and progressed into the ventral pathway. L, left; R, right.

These data suggest that a spatio-temporal junction exists between the occipital and occipito-ventral cortex around 170 ms, after which only behaviorally relevant features flow into the temporal cortex, with the processing of irrelevant features ending in the occipital cortex. In the next two sections, I detail what happens before and after this junction.

4.2.5.6 Dynamic Reduction of Nondiagnostic Brain Features in the Occipito-Ventral Pathway

I now investigate the temporal and spatial properties of diagnostic and nondiagnostic features representation in the occipital cortex towards the occipito-ventral junction, where they diverge ~170 ms post stimulus. To do this, I used each observer's 3D representation matrix and computed the maximum representation strength (i.e., MI effect size) across nondiagnostic (vs. diagnostic) brain features, separately for each voxel and time point. Specifically, for each observer I proceeded in two steps:

Step 1: Dynamics of brain feature representation between 0 and 400ms post stimulus. For each participant, I used their representation matrix and selected the voxels with significant MI for at least one nondiagnostic brain feature in the 0 to 400ms time window (henceforth, "nondiagnostic voxels"). For each nondiagnostic voxel, at each time point, I extracted the maximum MI over all nondiagnostic brain features to plot the maximum representation curve of this voxel. Figure 4-3A shows the representation curves of all nondiagnostic voxels. The curve of each voxel had an *onset* (the first time point at which maximum MI was significant) and an *offset* (the last time point of significance) that I computed; representation *duration* on a voxel was therefore computed as offset - onset. To capture the spatial properties of nondiagnostic voxels, I further computed the *Euclidean distance* (in the common MNI space) of each voxel in relation to the voxel with the earliest onset. I repeated above computations separately for diagnostic brain features.

For nondiagnostic voxels (vs. diagnostic voxels), I fitted a robust linear regression line between their onset times and Euclidean distances from the voxel of initial onset. I computed another robust linear regression between their representation duration and Euclidean distances (see panel A and B in Figure S4-4 to S4-8 for individual results in ***4.4.3*** *Supplemental Figures*). Table S4-2 and S4-3 detail the statistics of the robust linear regressions (see ***4.4.2*** *Supplemental*

Tables). I excluded outlier voxels for these analyses—i.e., voxels with > 3 standard deviations from the median onset of all voxels, computed separately using nondiagnostic and diagnostic voxel onset distributions, see Table S4-4 for the percentage of voxels exclusion in ***4.4.2*** *Supplemental Tables*.

Step 2: Spatial-temporal junction of divergence between nondiagnostic and diagnostic feature representations. I selected the voxels representing nondiagnostic features that were furthest in the brain--i.e., with Euclidean distances > 75th percentile of distances of all nondiagnostic voxels. These voxels represented the spatial marker of the junction. I defined the latest representation offsets of these voxels as the temporal marker of the junction (see Figure 4-3A the vertical dash line on the representation curves). To identify the brain regions (based on the "Talairach Demon Atlas" warped into MNI space) involved in the junction, I grouped nondiagnostic voxels of each observer according to their location in the cuneus (CU), lingual gyrus (LG), inferior occipital cortex (IOG), middle occipital gyrus (MOG), superior occipital gyrus (SOG), fusiform gyrus voxels locates quite close to LG (LG/FG, see Figure S4-16 for location), fusiform gyrus (FG), inferior temporal gyrus (ITG), middle temporal gyrus (MTG), superior temporal gyrus (STG), inferior parietal lobe (IPL), and superior parietal lobe (SPL). In each anatomical region, I then checked the Euclidean distance (see step 1) of all nondiagnostic and diagnostic voxels (see panel C of Figure S4-4 to S4-8 for individual results in ***4.4.3*** *Supplemental Figures*).

Figure 4-3A shows the representation time courses and brain scatters, which illustrates the dynamic reduction of nondiagnostic feature representations in each participant. Specifically, nondiagnostic feature representations initially travel as a wavefront that then reduces in duration as it progresses through the occipital cortex (c.f. the linear regression between Euclidean distance and duration in Figure S4-4A to S4-8A and Table S4-2 in ***4.4*** *Supplemental Materials*). Thus, the wavefront of nondiagnostic feature representations rapidly collapses (around 170 ms) as it travels into the occipital cortex. In contrast, identical computations applied to diagnostic features (see Figure 4-3B) demonstrate that the diagnostic wavefront progresses past 170 ms and deeper into ventral and dorsal regions. Figure S4-3C summarizes the anatomical brain regions where the two wavefronts diverge (see Figure S4-4 to S4-8 for each observer in ***4.4.3*** Supplemental Figures).

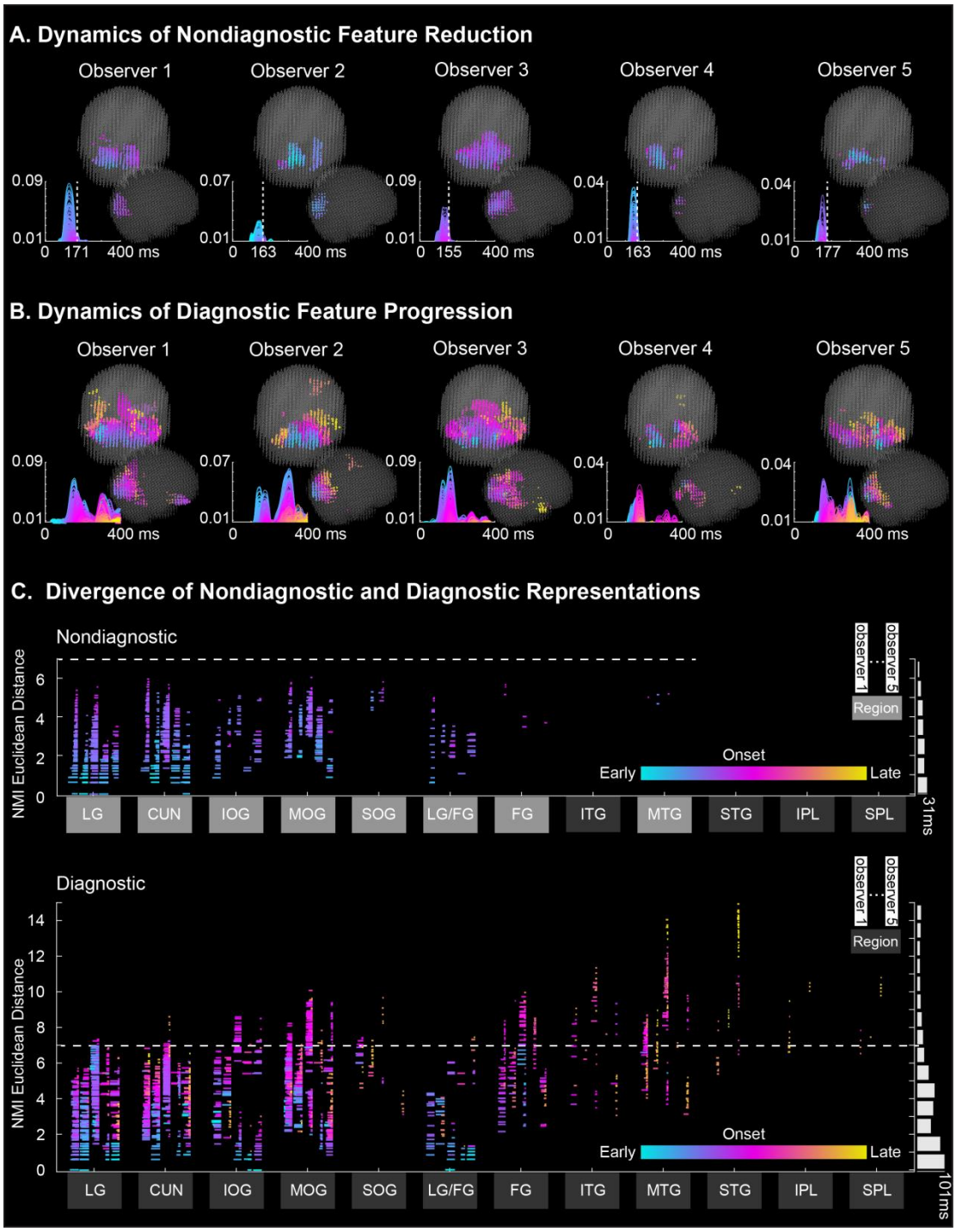

Figure 4–3 Dynamic Reduction of Nondiagnostic Brain Features in the Occipital-Ventral Pathway. Dynamics of (A) Nondiagnostic Brain Feature Reduction and (B) Diagnostic Brain Feature Progression. For each observer, a plot shows the curves of maximum (A) nondiagnostic and (B) diagnostic brain feature representation (i.e. MI effect size) for each voxel between 0 and 400ms post stimulus, color-coded by ranked onset time (blue, early; magenta, midway; yellow, late). In (A), the vertical dashed lines represent the time (~170 ms) at which the brain stops representing nondiagnostic features. Adjacent brain scatters locate the voxels associated with each curve using the same color code. C. Divergence of Nondiagnostic and Diagnostic Feature Representations. In each panel, brain regions comprise one column per observer, where each horizontal line represents one voxel from the region. Lines denote two voxel properties: the color denotes representation onset, and the length, representation duration.

Adjacent white bars show median representation duration across all regions, organized by the Y-axis of MNI Euclidean distance of each voxel to the voxel of initial representation onset. The dashed white horizontal line shows the nondiagnostic wavefront extends ventrally in the LG up to the junction with the TG and FG, and dorsally with IPL and SPL (see regions shaded a lighter grey). The diagnostic wavefront extends further into the ventral (i.e. FG, ITG, MTG, and STG) and dorsal (i.e. IPL and SPL) (see pink to yellow colors). Abbreviations: Cuneus (CU), lingual gyrus (LG), inferior occipital cortex (IOG), middle occipital gyrus (MOG), superior occipital gyrus (SOG), fusiform gyrus (FG), inferior temporal gyrus (ITG), middle temporal gyrus (MTG), superior temporal gyrus (STG), inferior parietal lobe (IPL), and superior parietal lobe (SPL). Observer = participant.

4.2.5.7 Dynamic Construction of Behavior Representations in the Right Fusiform Gyrus:

Above results show that only diagnostic brain features are represented past the occipito-ventral, 170 ms junction. A prevalent hypothesis is that visual information represented early and separately across the left and right occipital cortices (*Niemeier, Goltz, Kuchinad, Tweed, & Vilis, 2005)* later converges in the rFG to support visual cognition tasks, such as visual decisions *(Ince et al., 2015*). However, conclusive testing of this hypothesis remains challenging for two reasons. First, the hypothesis implies the need to characterize the evolution of increasingly complex (e.g., lateralized to bilateral) stimulus representations in the dynamic brain activity of this specific region, and not others. Second, it requires demonstrating that the representations specifically support task behaviors. I now settle these 2 points step by step:

Step 1: Redundancy Computation. The CBBI framework introduces the calculation of feature redundancy (Red, *Ince et al., 2017*). Red quantifies the shared variability between: <Information Samples; MEG Activity; Perceptual Decision> on individual trials. It therefore directly measures modulations of feature representations in the brain to support each perception specifically.

Specifically, for each observer, we computed the triple relationship between <Brain Feature Coefficients; MEG Voxel Activity; “the nuns,” “Voltaire,” “don’t know”>:

$$Red = MI < Feature; Perceptual\ Decision > + MI < Feature; MEG\ Voxel\ Activity > - MI < Feature; MEG\ Voxel\ Activity, Perceptual\ Decision > \quad (5)$$

(5) is equivalent to the set theoretic intersection of three variable entropies, or alternatively the intersection of any two mutual information. I applied Equation

(1) for each combination of diagnostic brain feature, brain voxel, and every 2 ms between 0 and 400 ms post stimulus. This computation produced a 3D redundancy matrix (feature × voxel × time point). I established statistical significance for each cell by recomputing redundancy with shuffled decision responses across trials (repeated 200 times), and used the 95th percentile of 200 maximum values (each taken across of the entire 3D redundancy matrix per permutation) as statistical threshold (i.e. FWER, $p < 0.05$, one-tailed).

Step 2: Representational Complexity Computation. If information converges on a brain region to support task behavior, then the number of features represented in the region's voxels should increase over time - an increase in the complexity of the region's population code. For each participant, I quantified representational complexity for behavior by counting the number of different features that each brain voxel represents redundantly with behavior, independently in 5 evenly distributed time intervals over the between 120-220 ms post stimulus (see grey level scatters in Figure S4-9A to S4-13A for individual participant in ***4.4.3*** *Supplemental Figures*). This specific time interval encompasses the M/N170 time course (*Bentin et al., 1996; Schyns et al., 2007*). To represent the complexity at the group level, in each time window and for each of the 12,773 brain voxels, I calculated the median number of different redundant features it significantly represented across five participants.

As shown in Figure 4-4A, representational complexity does indeed increase over time and peaks between 161 – 201ms, primarily in the rFG (see also Figure S4-9A to S4-13A for this increase in each participant in ***4.4.3*** *Supplemental Figures*).

Step 3: Representation of Behavior in the Brain. To confirm the behavioral relevance of such complexity, for each voxel, I also computed MI <MEG Voxel activity, "the nuns," "Voltaire," "don't know" >, resulting in a 2D voxel by time matrix. To establish statistical significance, I extracted the maximum MI across the matrix recomputed, shuffling decision responses across trials in each cell (repeated 200 times). I used the 95th percentile of this maximum distribution as statistical threshold (i.e., FWER, $p < 0.05$, one-tailed). Figure 4-4B shows the behavioral representation in FG voxels at group level in each time window, using the median of MI values (max in each window) cross 5 participants (see also orange scatters in Figure S4-5.1B to S4-5.5B for individual participant in ***4.4.3***.

Supplemental Figures). As demonstrated, the perceptual decision peaks in brain activity in the time window during which representation complexity peaks (comparing Figure 4-4A and B).

Step 4: Decision-specific Feature Representations. I further decomposed representational complexity into the specific features that underlie each perceptual behavior, in each individual observer. Specifically, I uncovered the perception-specific redundant features of each observer by computing information theoretic redundancy between <Brain Feature Coefficients; MEG Voxel Activity; "the nuns," "don't know">, and between <Brain Feature Coefficients; MEG Voxel Activity; "Voltaire," "don't know">, separately. I used the permutation test described in Step 1 to threshold redundancy values and obtain the features represented on rFG voxels for each perceptual decision behavior (see color-coded scatters in Figure. S4-9 C to S4-13 C for each participant in ***4.4.3****. Supplemental Figures*).

Figure 4-4C shows the perception-specific redundant features representation in the fourth time window, when both the feature complexity and behavioral representation peak (see Step 2 and 3). As shown, voxels at the top of the rFG represent redundant features that are linked to the response "Voltaire" (primarily the green global face in SF3, the right orange eye in SF1, and the right red eye in SF2). Other redundant features are primarily linked to the response "the nuns" (the turquoise left face in SF1, and the blue and red faces in SF2). Note also that the representation of ipsi-lateral information in the rFG (e.g., the orange and red features) implies that inter-hemispheric information transfer occurs from its initial contra-lateral representation in the left occipital cortex (see Figure 4-1C and *Ince et al., 2015*).

Thus, by using feature redundancy and representational complexity, I have demonstrated that rFG voxels represent stimulus information with a selectivity and complexity that supports task-specific behaviors.

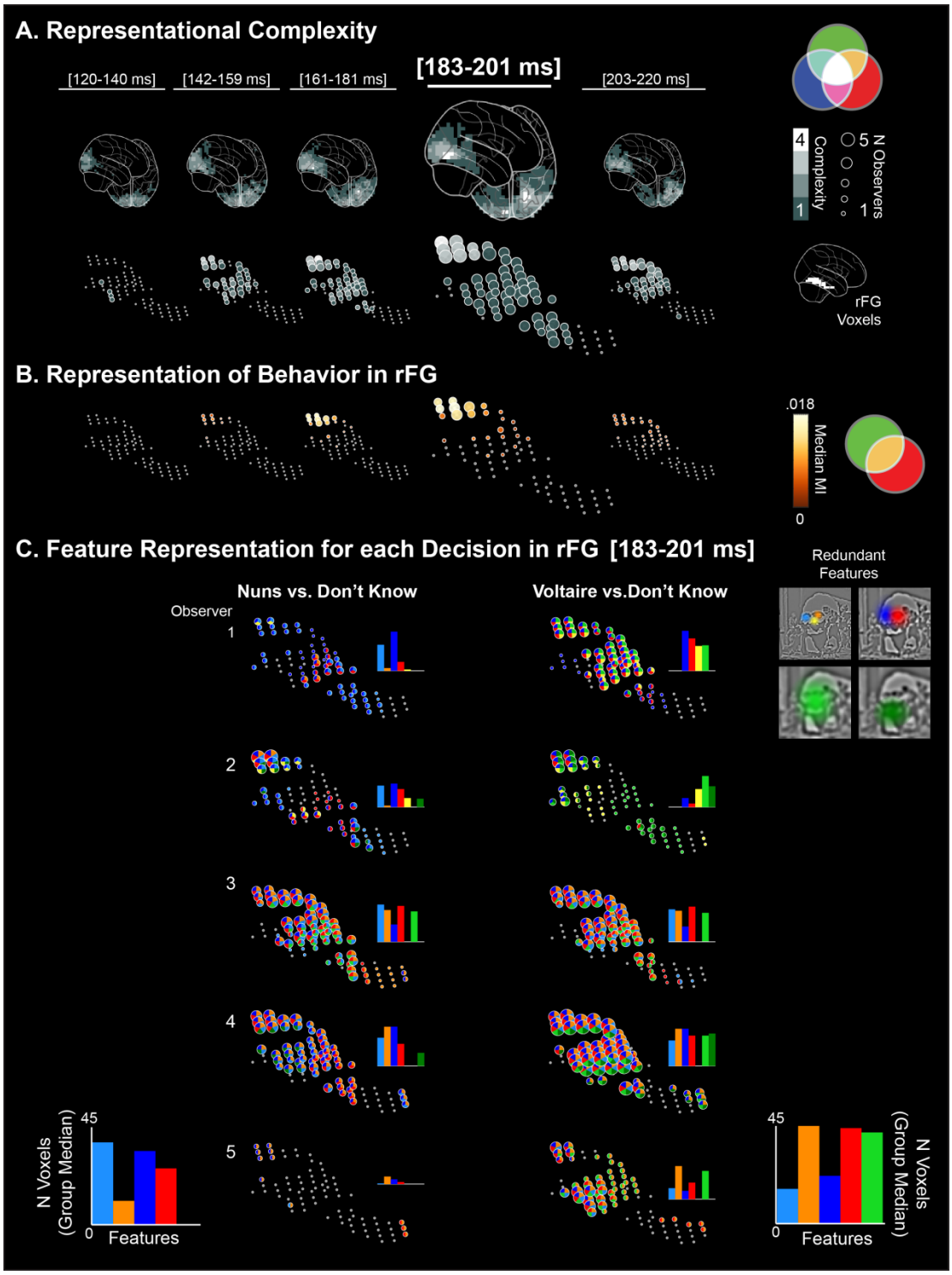

Figure 4–4 Dynamic Construction of Behavior Representations. A. Representational Complexity. Grey level voxels in each brain schematic and in each time window denote the median number of redundant behavioral features represented across observers. Times in brackets indicate the range of each time interval (time started and ended). Beneath, voxels in the rFG show that representational complexity peaks at the top of the rFG in the fourth [183 – 201 ms] time window (highlighted). Voxel size denotes the number (N) of observers who represented at least one redundant behavioral feature on this voxel and time window. B. Representation of Behavior. Yellow voxels in each time window denote the median MI between MEG activity and the decisions "the Nuns," "Voltaire," "don't know" across observers (illustrated with the yellow intersection in the Venn diagram). C. Feature Representation for each Decision in rFG.

Representational complexity was decomposed at each rFG voxel and time window by showing features that are redundantly represented in MEG activity and for each behavioral decision, in each observer (see adjacent color-coded features). Adjacent histograms show the number of rFG voxels representing each redundant feature. The bottom histograms show the median number of voxels representing each redundant feature across observers, showing feature selectivity for each decision (e.g., the turquoise HSF left nuns face and the green LSF bust of Voltaire).

4.3 Discussion

In this case study, I investigated how high-dimensional information input collapses in the occipito-ventral pathway to become low dimensional representations that guide behavior, using a novel information theoretic framework called CBBI. Using this framework, I identify that high dimensional stimuli are reduced in the occipito-temporal pathway into low dimensional representations that can support subsequent perceptual decision making. To address the where, when and how of information processing, I tracked dynamic feature representations in the brain and show that behaviorally irrelevant information is rapidly reduced at the occipito-ventral junction around 170ms. The results also showed that rFG representations for behavior are constructed between 161 and 201 ms, post stimulus. Remarkably, I replicated all these results independently in 5 individual observers, as is now better practice, with high effect sizes, in part due to high trial numbers. Specifically, using non-parametric family-wise error rate correction I found spatio-temporally coincident significant effects within all five observers. This is a substantially stronger finding than conventional cluster corrected group statistics, where the effects can be driven by a small subset of participants and are usually non-significant within any individual observer. Thus, CBBI enabled us to interpret the information processing of task-related brain activity because it computes the interactions between three variables in individual observers, rather than two across groups of observers, as is typical in neuroscience and neuroimaging – i.e. either the interaction between stimulus and neural response, or stimulus and behavior, or neural response and behavior, separately. By directly computing the interactions between all three variables (i.e. the three double interactions plus the full triple interaction, cf. the colored set intersections) CBBI addressed the recent argument (*Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017*) that neuroscientific explanations need to include behavior to better tease apart the component processes of the brain.

4.3.1 Information Reduction in the Occipito-Ventral Pathway

The information reduction process I documented evolves over time from a state of many to a state of fewer dimensions of stimulus representation. Such many-to-fewer transition does not imply that the transition involves only feedforward processes. Instead, the hierarchical layers of the occipito-ventral pathway likely communicate with each other using both feedforward and feedback signals to implement the data reduction over time. Such interactive architectures of are similar to well-known network models that resolve ambiguity between hierarchically organized representations (*T. S. Lee & Mumford, 2003; Mcclelland & Rumelhart, 1981*). I subscribe to this interactive organization whereby selection of diagnostic features from the visual stimulus probably involves memory predictions, which propagate down the visual hierarchy and interact with the feedforward flow (*Bar, 2007; K. Friston, 2008; Slotnick & Schacter, 2004; M. L. Smith, Gosselin, & Schyns, 2012*). Although in this study I can visualize the feedforward flow of stimulus representation by coupling information samples with subsequent brain responses, the arrow of time prevents me from visualizing the representation of top-down predictions in brain activity (though see *M. L. Smith et al., 2012* for such visualizations from behavior). Nevertheless, this interactive architecture could be further documented by visualizing successive transformations of stimulus representations over time.

I traced the dynamic representation of a nun's face (the HSF pixels representing this image feature) from occipital cortex into the ventral pathway. It would be naïve to assume that the nun's face is represented as such in any of these regions, but we need a broad view of the information-processing, which this approach affords. To better understand the transformation of representations along the visual hierarchy, researchers would need to sample an explicit hierarchical generative model of visual information of face, objects and scenes (with invariant representations at the top of its hierarchy and with Gabor-type filters at the bottom, which better model stimulus feature representations in the early visual cortex *(Kay, Naselaris, Prenger, & Gallant, 2008)*. Likewise, models tolerant to changes in size, rotation, and illumination would better reflect properties of higher-level ventral pathway representations. Designing such generative models remains a necessary but considerable challenge to understanding complex sensory representations (see also *Olman & Kersten, 2004; Ponsot, Burred, Belin, & Aucouturier, 2018; Zhan, Garrod, Van Rijsbergen, & Schyns, 2017; Zhu, 2007*).

4.3.2 Time Course of Information Processing in the Occipito-Ventral Pathway

The information processes at the occipito-ventral junction flank the timing and sources of the N/M170 event related potentials *(Horovitz, Rossion, Skudlarski, & Gore, 2004)*, which reflects a network that represents and transfers features across the two hemispheres *(Ince et al., 2016)*. I showed earlier that the N/M170 first represents the contra-lateral diagnostic eye prior to the N170 peak, followed by the ipsi-lateral diagnostic eye, transferred after the N170 peak from the opposite hemisphere. The timing of this process is analogous to that documented here in Fig. 4. Potentially, the N170 peak might reflect the event when the wavefronts of behaviorally relevant and irrelevant information diverge. Alternatively, the pre- and post-170 ms rFG time courses could reflect two processing stages. Pre-170 ms, rFG could buffer visual information arising first from the contra-lateral visual field, followed by the ipsi-lateral visual field information that is transferred from the left occipital hemisphere; post-170 ms, rFG could integrate this buffered information across the two visual fields, as shown here. Our results should generalize to expert categories *(Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999*) and to the overlapped rFG representations of faces, objects and scenes (*Grill-Spector, 2003*). Future research can tease these apart within CBBI and characterize their category-specific representation dynamics in the occipito-ventral and dorsal pathways.

4.3.3 Relationship between Information Reduction in Occipital Cortex and Consolidation in rFG

Our CBBI results inform early vs. late attentional models of information selection *(Driver, 2001*), though we must be careful with interpretation because our study was not specifically designed to address them. We identified an occipito-ventral spatiotemporal junction that constrains where and when feature reduction occurs—i.e. in occipital cortex, before 170 ms—and where and when feature consolidation for perceptual decision occurs—i.e. in rFG, from 170 ms. Our data show that reduction involves other regions than V1-V2, though these could influence selection with gain control mechanisms (*Hillyard, Vogel, & Luck, 1998; Schwartz & Simoncelli, 2001*). However, reduction is not as late as rFG, because this region mainly represents diagnostic features. Thus, the spatiotemporal characteristics we report is akin to mixed model of attentional selection. CBBI

offers a powerful platform to directly study the attentional mechanisms for feature reduction and selection in complex tasks.

To conclude, the CBBI framework enables us to investigate task-sensitive brain activity that relates information processing in the brain to behavior. CBBI enables brain processes to be isolated (here, the reduction of behaviorally irrelevant information and the construction of behavioral representations), and employs principles that are broadly applicable across different modalities and granularities of brain measures used in a wide range of cognitive and systems neuroscience.

4.4 Supplemental Materials

4.4.1 Supplemental Method

*K-means of Brain Features**.*** Observers' brains represented similar brain features in the task (see Figure S4-1B). This warranted their projection onto a common feature basis for group-level visualization. To this aim, we applied k-means clustering by setting *k*, the number of clusters, to 25, to align the number of means to the maximum number of NNMF brain features computed in any observer. We pooled the normalized NNMF brain features of all observers, resulting in a 115 x 20480 matrix (115 NNMF components in total for 5 observers and 64 pixels * 64 pixels * 5 SFs weights). We applied k-means (cosine similarity, 1000 repetitions) to this matrix (see Figure S4-15 for the resulting k-means clusters). It is important to emphasize that we performed all analyses on the specific brain features of each observer. I only indexed these individual features onto the common k-mean feature basis and corresponding color codes to report group results (e.g. in Figure 4-1 and 4-4). See Figure S4-6D *K-means of Brain Features* for a graphic illustration of the process.

4.4.2 Supplemental Tables

Table S 4-1 Number of trials following pre-processing of MEG data.

Participant	All responses	"Nuns" response	"Voltaire" response	"Don't Know" response
1	3314	1189	1313	812
2	3604	1666	1263	675
3	4154	1634	1892	628
4	3023	1603	738	682
5	2885	1007	1346	532

Table S 4-2 Nondiagnostic voxels. Linear models between the Euclidean distance (Y) and Onset/Duration (X), and p values for the slope, per observer.

	Onset		Duration	
Participant	model	p value	model	p value
1	Y=0.118X - 3.546	p < .001	Y = -0.063X + 4.906	p < .001
2	Y=0.094X - 3.222	p < .001	Y = -0.057X + 3.673	p < .001
3	Y=0.084X -0.457	p < .001	Y = -0.064X + 5.189	p < .001
4	Y=0.201X - 10.535	p < .001	Y = -0.083X + 3.746	p < .001
5	Y=0.154X - 7.912	p < .001	Y = -0.061X + 2.869	p < .001

Table S 4-3 Diagnostic voxels. Linear models between the Euclidean distance (Y) and Onset/Duration (X), and p values for the slope, per observer

	Onset		Duration	
Participant	model	p value	model	p value
1	Y=0.024X + 3.214	p < .001	Y = -0.025X + 7.013	p < .001
2	Y=0.030X + 1.343	p < .001	Y = -0.021X + 5.852	p < .001
3	Y=0.055X + 2.462	p < .001	Y = -0.015X + 7.408	p < .001
4	Y=0.174X - 8.126	p < .001	Y = -0.011X + 5.004	p = .094
5	Y=0.000X + 4.061	p = .919	Y = -0.014X + 4.693	p < .001

Table S 4-4 Percentage of voxels excluded from onset analysis.

Participant	Diagnostic	Nondiagnostic
1	1.81%	0.35%
2	0	2.45%
3	2.21%	0.19%
4	3.52%	0
5	0.58%	0

4.4.3 Supplemental Figures

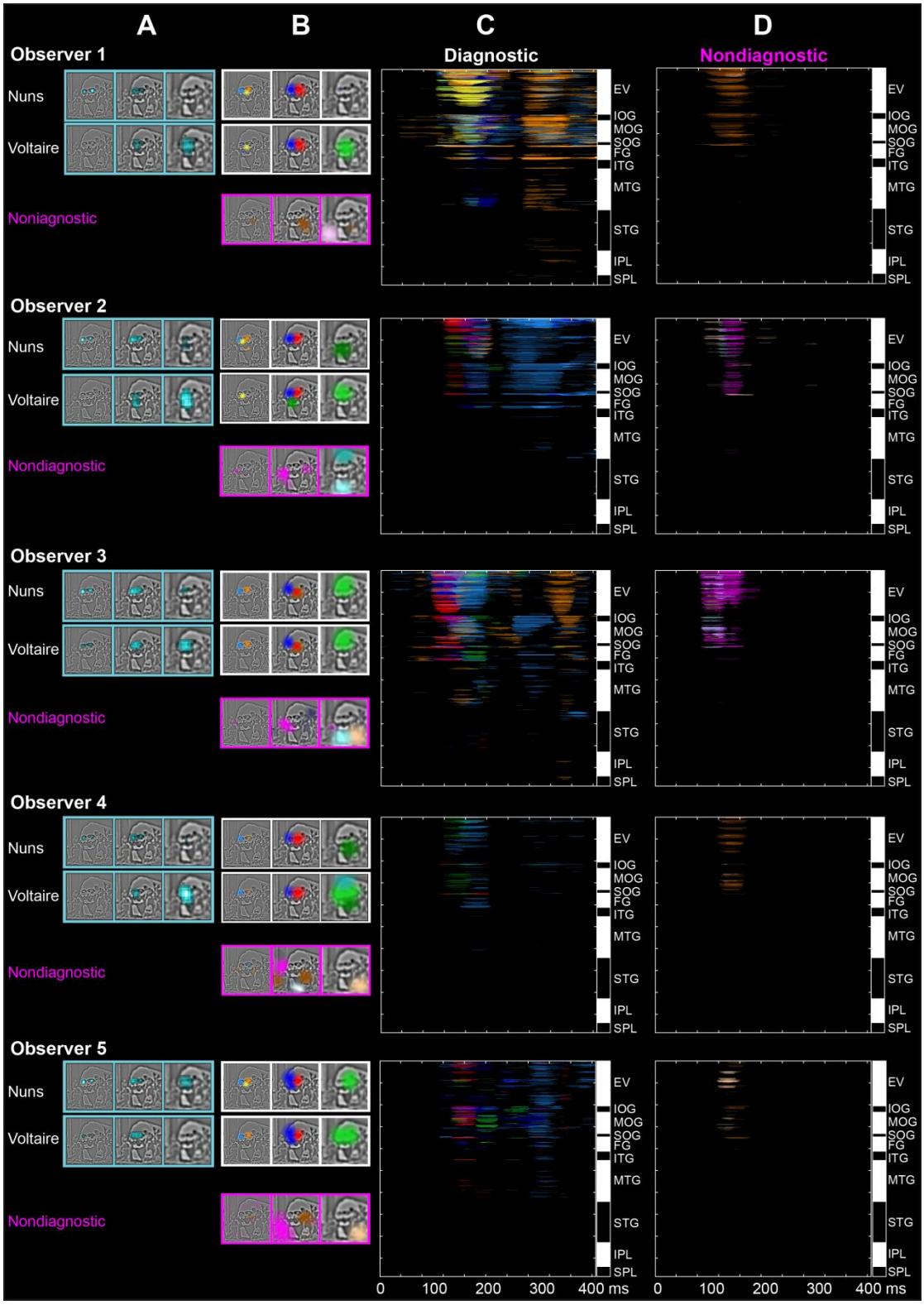

Figure S 4-1 Diagnostic Features and Brain Features of each Observer. Diagnostic Features. The cyan framed images in column A report the significant (FWER $p < 0.01$, one-tailed) MI value for each pixel in the first three SF bands, revealing across observers the features most diagnostic for responding "the nuns" (the two small faces in SF band 1) and "Voltaire" (the broad face in SF band 3). Brain Features. White frames in column B highlight the diagnostic features that MEG voxels represented, separately presented for decisions "the nuns" and

“Voltaire;” Magenta frames in column B highlight the nondiagnostic features that MEG voxels represented. In column C and D, each cell of the Diagnostic (column C) and Nondiagnostic (column D) representation matrices report the color-coded significant brain feature with maximum representation in MEG effect size (i.e. MI) across all brain features, at this voxel and time point. For reference, alternating white/black bars flanking each matrix indicate the anatomical brain region of the corresponding voxels. To illustrate, the representation matrices of Observer 1 reveal that the diagnostic brain feature “nose of Voltaire” in yellow is primarily represented in specific EV, MOG and FG voxels with highest effect sizes across the full-time course. In contrast, the brown nondiagnostic brain feature is primarily represented in occipital regions, and before ~170 ms. EV = early visual cortex (including lingual gyrus and cuneus), IOG = inferior occipital gyrus, MOG = middle occipital gyrus, SOG = superior occipital gyrus, FG = fusiform gyrus, ITG = inferior temporal gyrus, MTG = middle temporal gyrus, STG = superior temporal gyrus, IPL = inferior parietal lobe, SPL = superior parietal lobe.

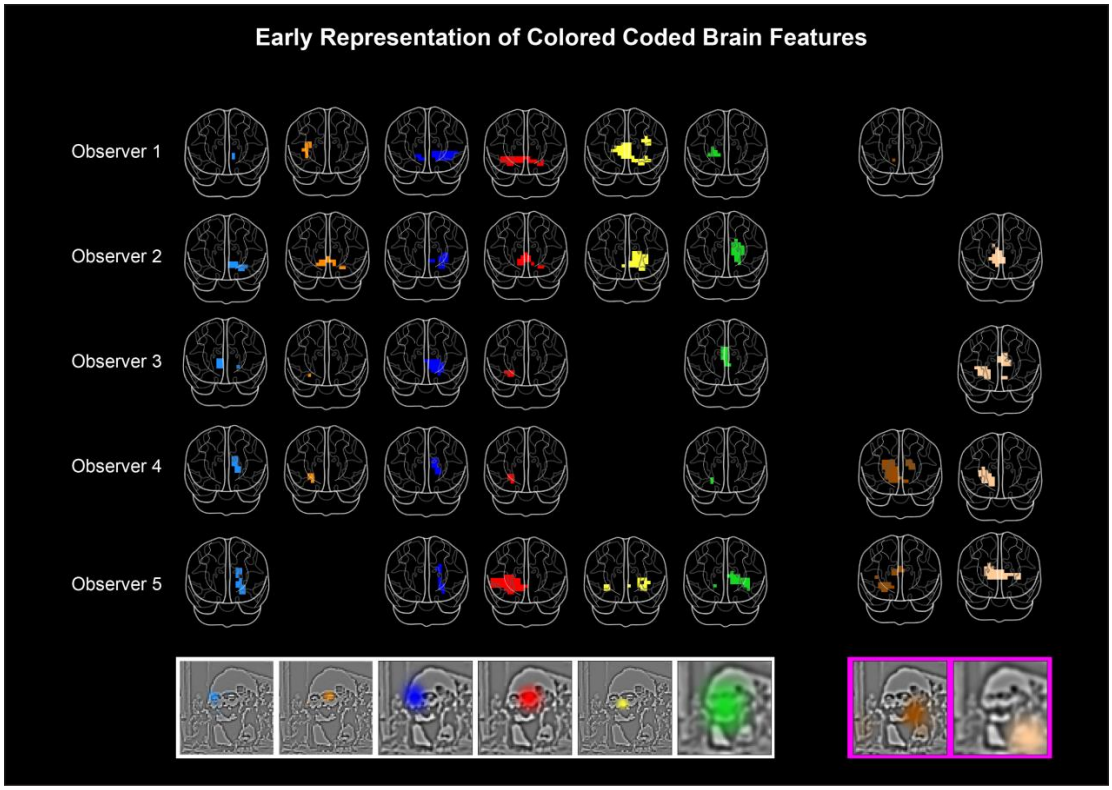

Figure S 4-2 Early Representation of Brain Features. For each observer, color-coded brain regions show the early (during the initial 20 ms of representation) topological representation of diagnostic and nondiagnostic brain features (FWER, $p < 0.05$, one-tailed). Here, we only show the common brain features represented by the majority of observers (see Figure S4-1B for the representation of all diagnostic and nondiagnostic features in each observer).

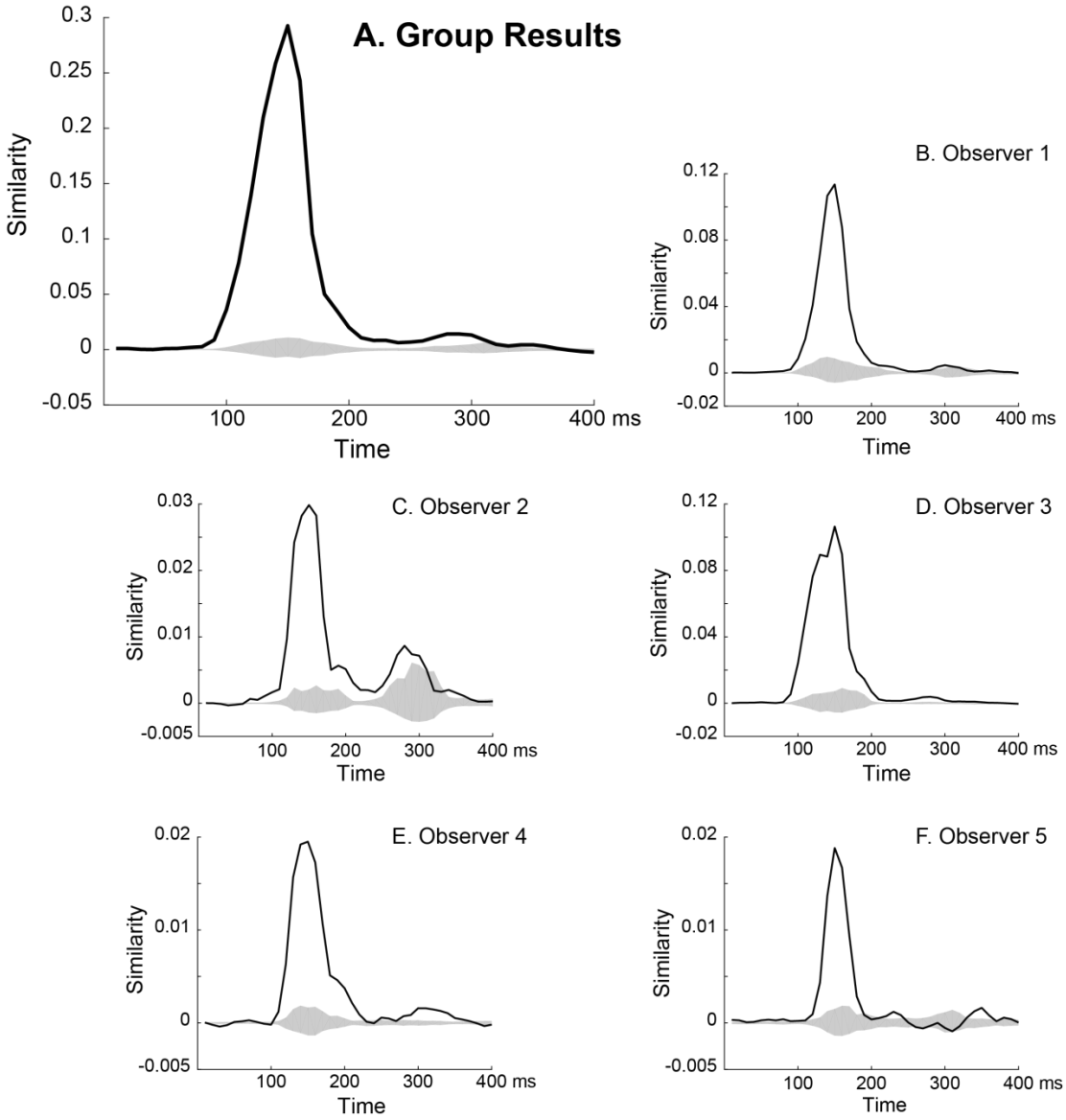

Figure S 4-3 Divergence of Diagnostic and Nondiagnostic Brain Features. A. Group Results. The plot shows the similarity (i.e. de-meaned dot product) between diagnostic and nondiagnostic brain feature representations over the time course of visual information processing. The shadowed region indicates the Bonferroni corrected chance-level similarity ($p < 0.05$, two-tailed). B to F. Diagnostic vs. nondiagnostic brain feature representation similarity for each observer. Together, the results show a consistent dynamic pattern of increasing similarity of diagnostic and nondiagnostic feature representations in the brain of each observer, up until 170 ms post-stimulus, following which diagnostic and nondiagnostic feature representations become dissimilar (i.e. diverge).

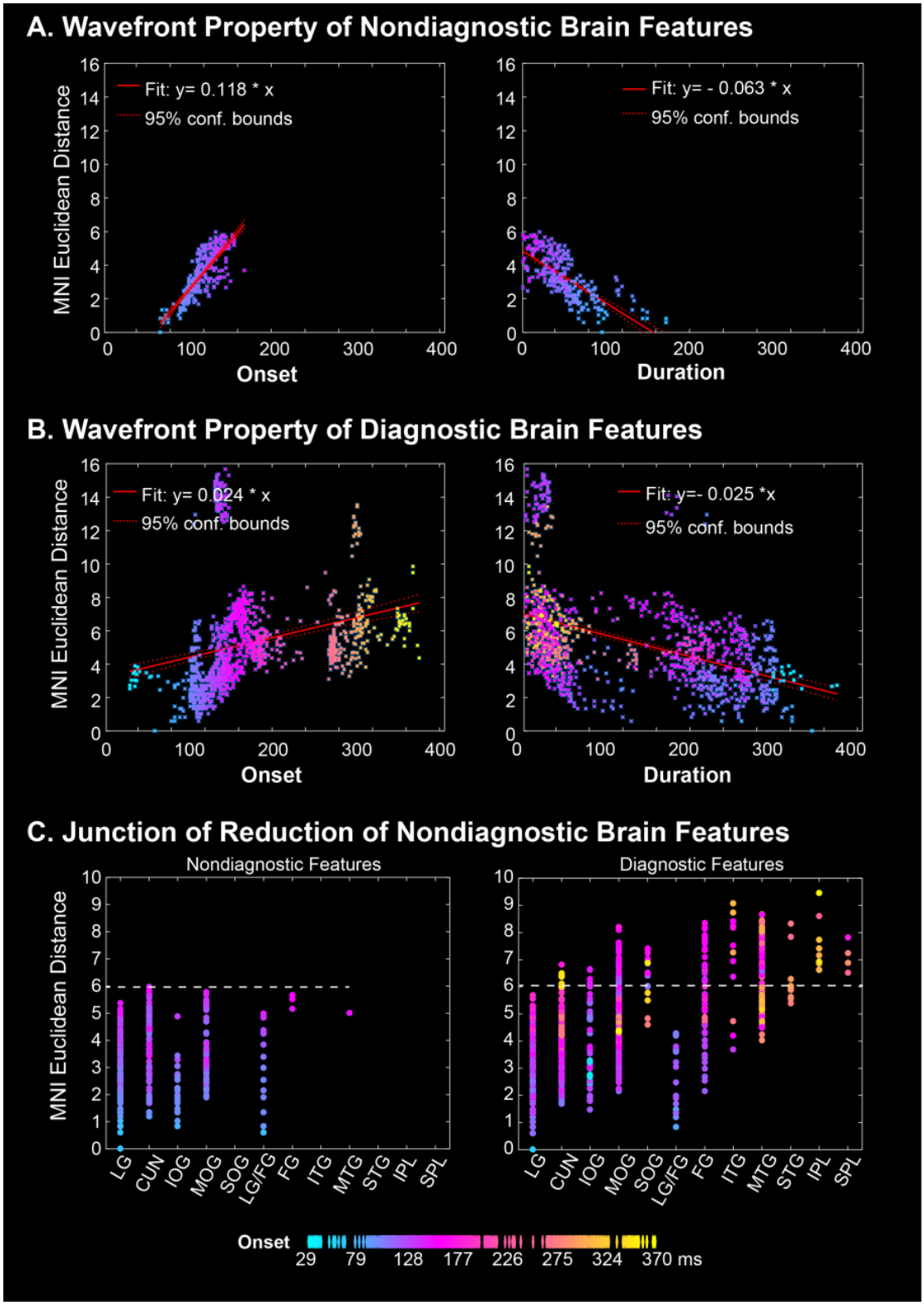

Figure S 4-4 Dynamic Reduction of Nondiagnostic Brain Feature Representations in Occipito-Ventral Pathway (Participant 1). A. Wavefront Property of Nondiagnostic Brain Feature Representations. The left scatter shows a linear relation between the representation onset times of voxels and their Euclidean distances to the voxel of initial representation onset. The right scatter shows that duration of nondiagnostic feature representation linearly decreases with the increasing distance of the considered voxel from the voxel of initial representation onset. B. Wavefront Property of Diagnostic Brain Feature Representation. Same caption as in panel A for diagnostic brain features, with later onsets (pink to yellow colors) in ventral and dorsal regions. C. Junction of

Reduction of Nondiagnostic Brain Features. In the left panel, voxels color-coded by onset times are pooled by anatomical brain region (X-axis) and scattered according to their Euclidean distance to the initial onset voxel of nondiagnostic representation on the Y-axis. In the right panel, the same caption for diagnostic voxels. The horizontal dashed line indicates the brain regions of furthest representation of nondiagnostic features. LG/FG on the X-axis comprises voxels located near to LG (see Figure S4-8 for location). LG = Lingual Gyrus, CUN = cuneus, IOG = inferior occipital gyrus, MOG = middle occipital gyrus, SOG = superior occipital gyrus, FG = fusiform gyrus, ITG = inferior temporal gyrus, MTG = middle temporal gyrus, STG = superior temporal gyrus, IPL = inferior parietal lobe, SPL = superior parietal lobe.

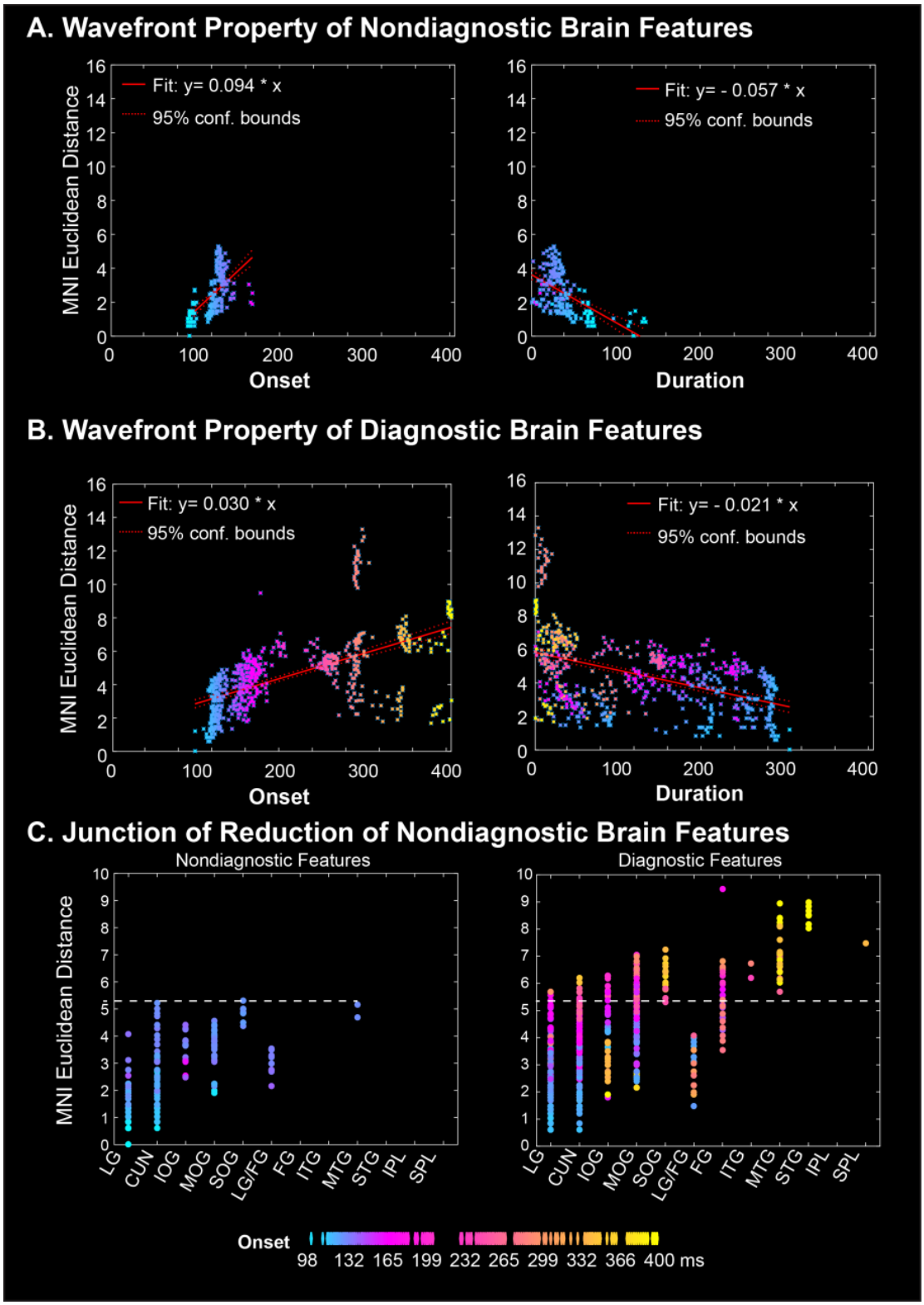

Figure S 4-5 Dynamic Reduction of Nondiagnostic Brain Feature Representations in Occipito-Ventral Pathway (Participant 2). Same caption as in Figure S4-4.

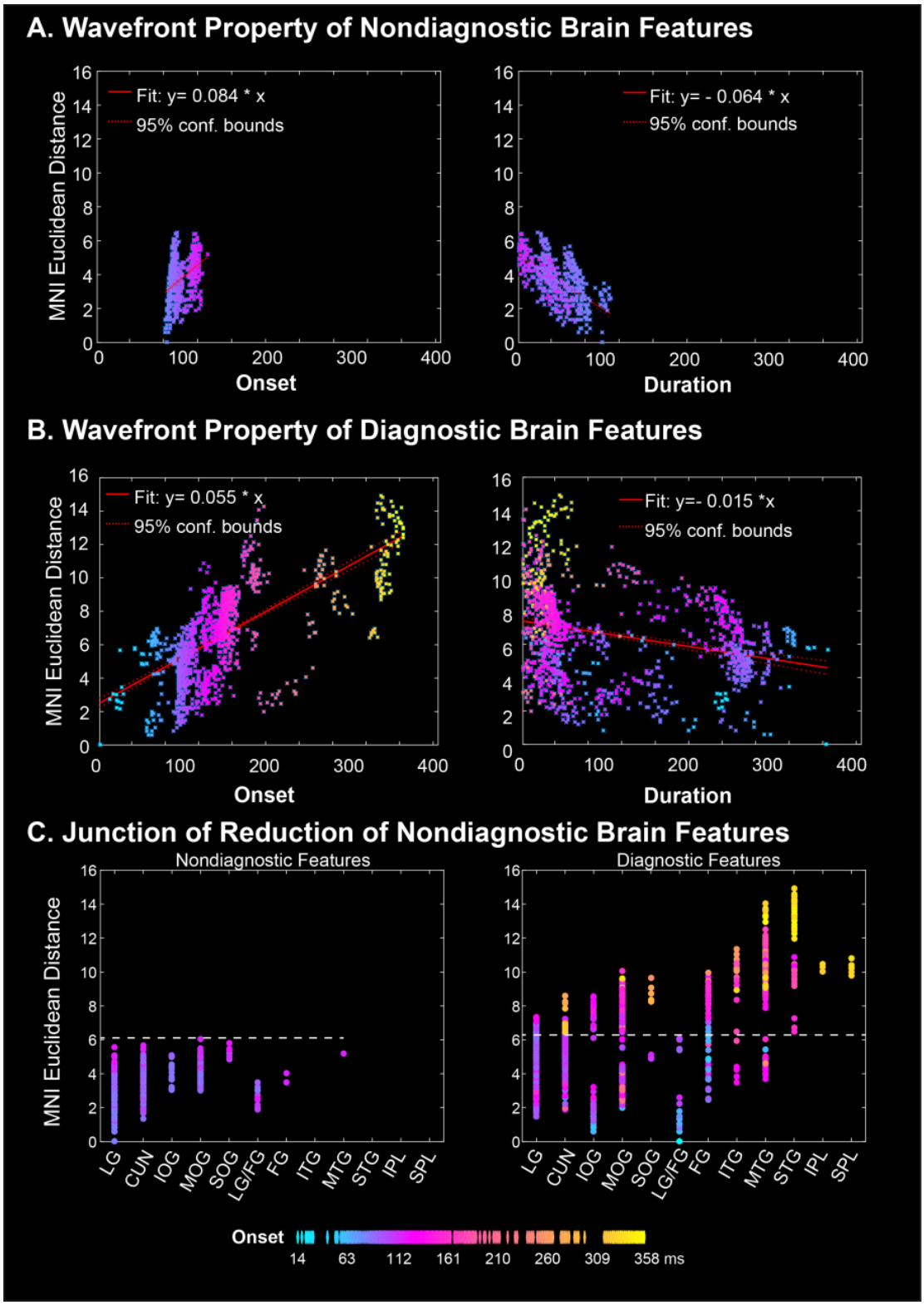

Figure S 4-6 **Dynamic Reduction of Nondiagnostic Brain Feature Representations in Occipito-Ventral Pathway (Participant 3).** Same caption as in Fig. S4-4.

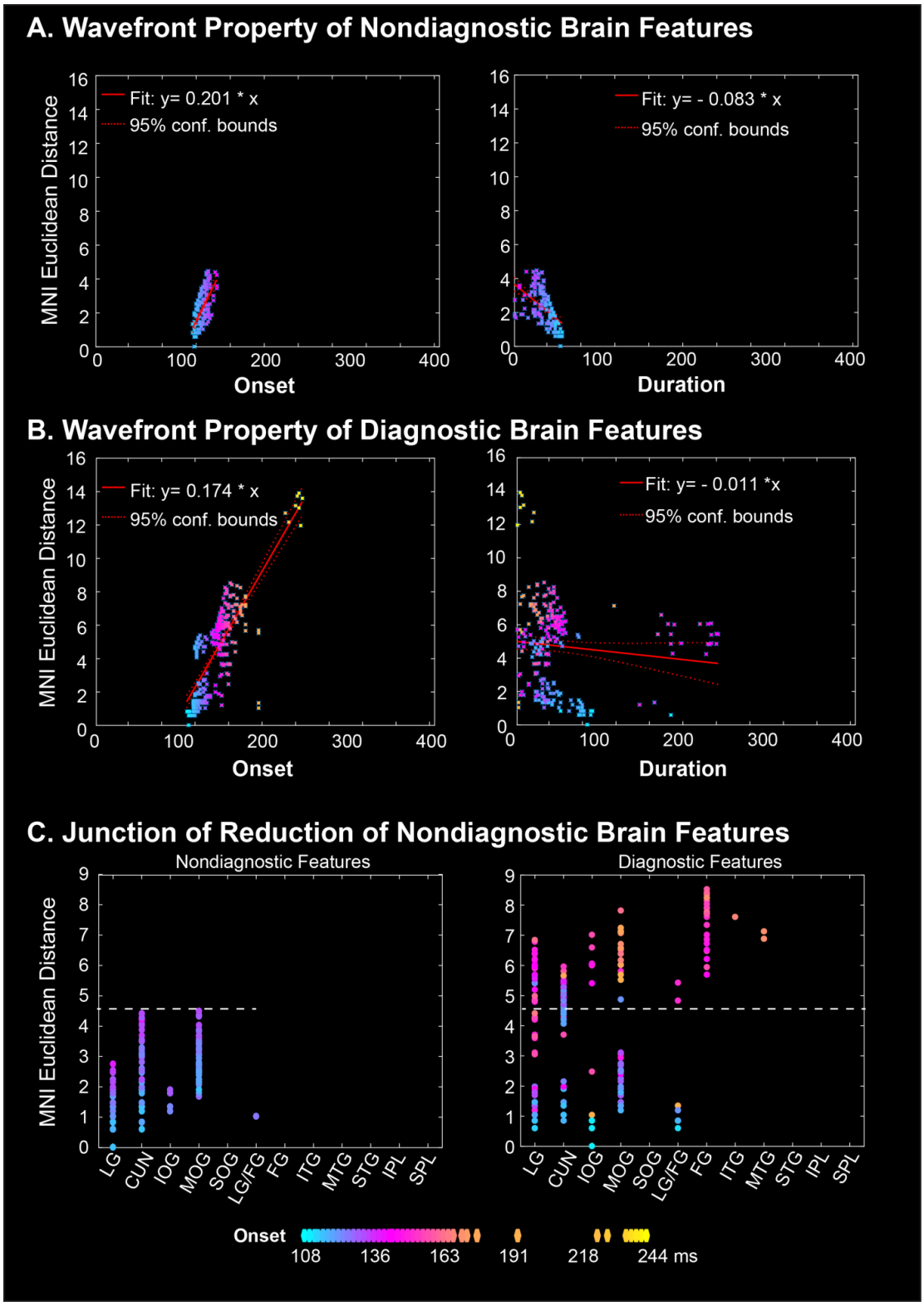

Figure S 4-7 Dynamic Reduction of Nondiagnostic Brain Feature Representations in Occipito-Ventral Pathway (Participant 4). Same caption as in Figure. S4-4.

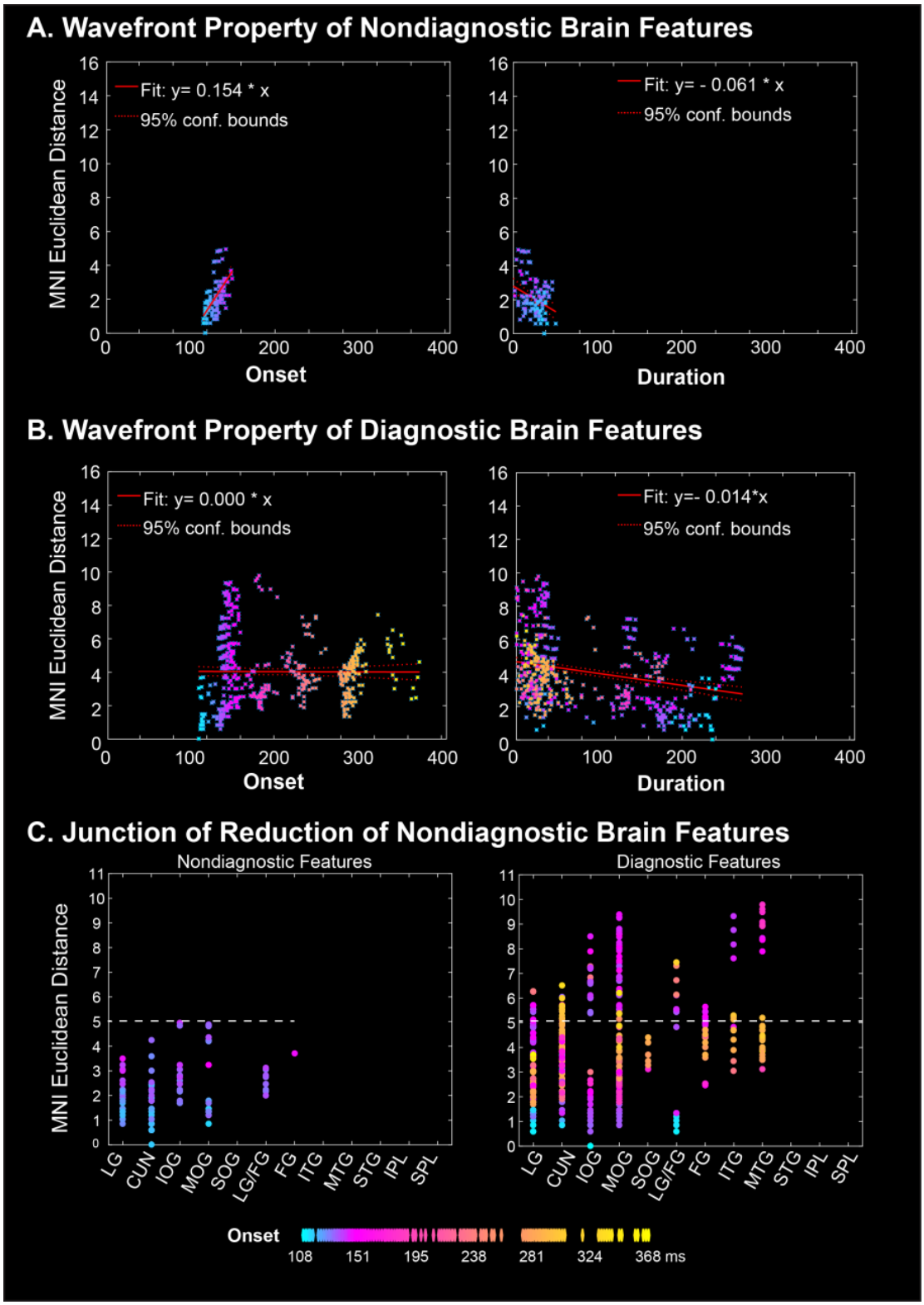

Figure S 4-8 Dynamic Reduction of Nondiagnostic Brain Feature Representations in Occipito-Ventral Pathway (Participant 5). Same caption as in Figure S4-4.

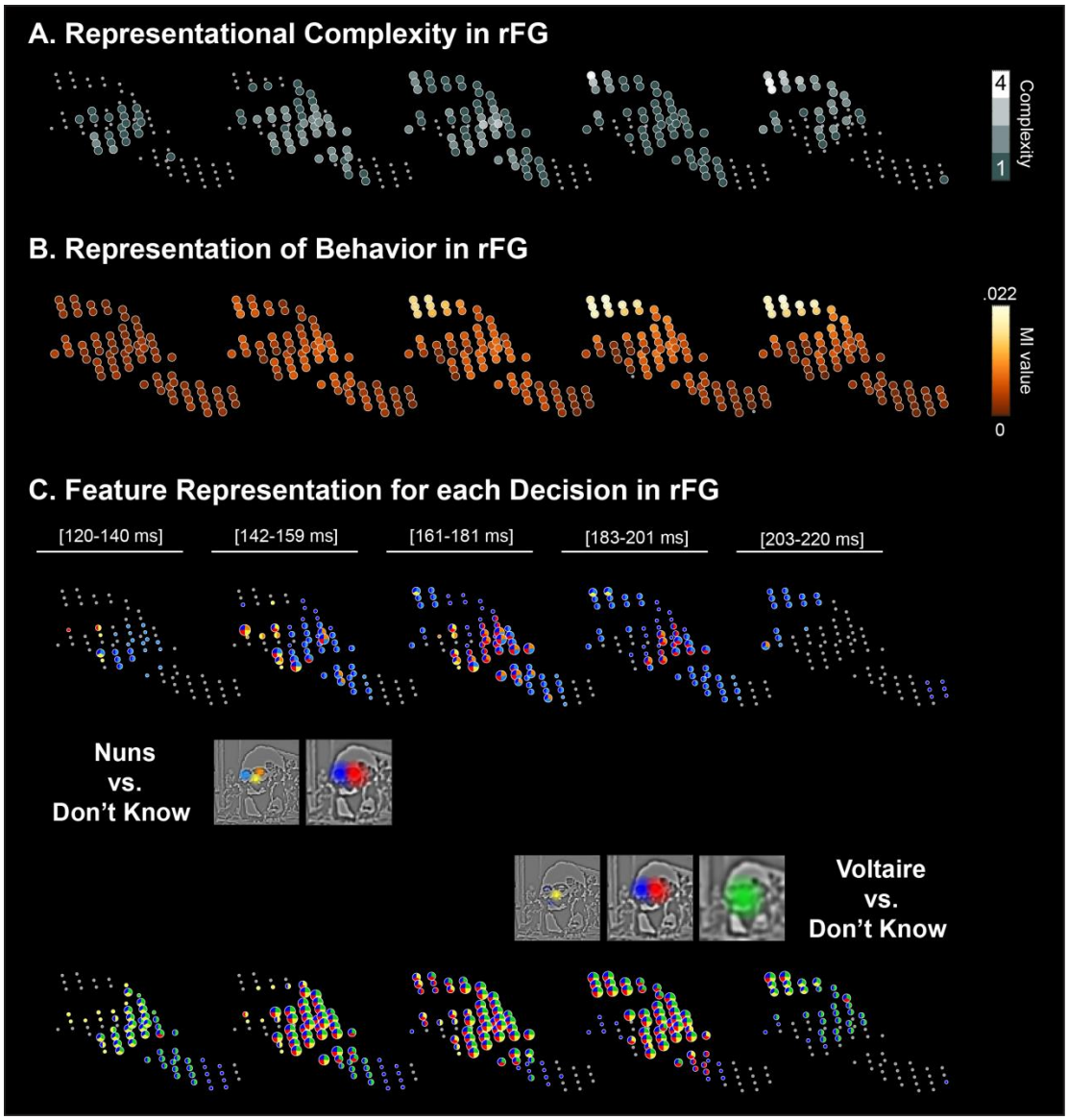

Figure S 4-9 Dynamic Construction of Representations for Behavior in rFG (Participant 1). A. Representational Complexity in rFG. Starting and ending times in brackets indicate the ranges of the time intervals for this observer. The grey level of the right Fusiform Gyrus (rFG) voxels corresponds to the number of redundant features that it represented within each time interval. B. Representation of Behavior in rFG. Yellow voxels denote the maximum MI (un-thresholded) between MEG activity and decisions "the Nuns", "Voltaire," "don't know" in each time interval. The yellow level represents the maximum MI value. C. Feature Representation for each Decision. Specific redundant features represented at each rFG voxel and time interval for each decision behavior (see the color-coded features for interpretation).

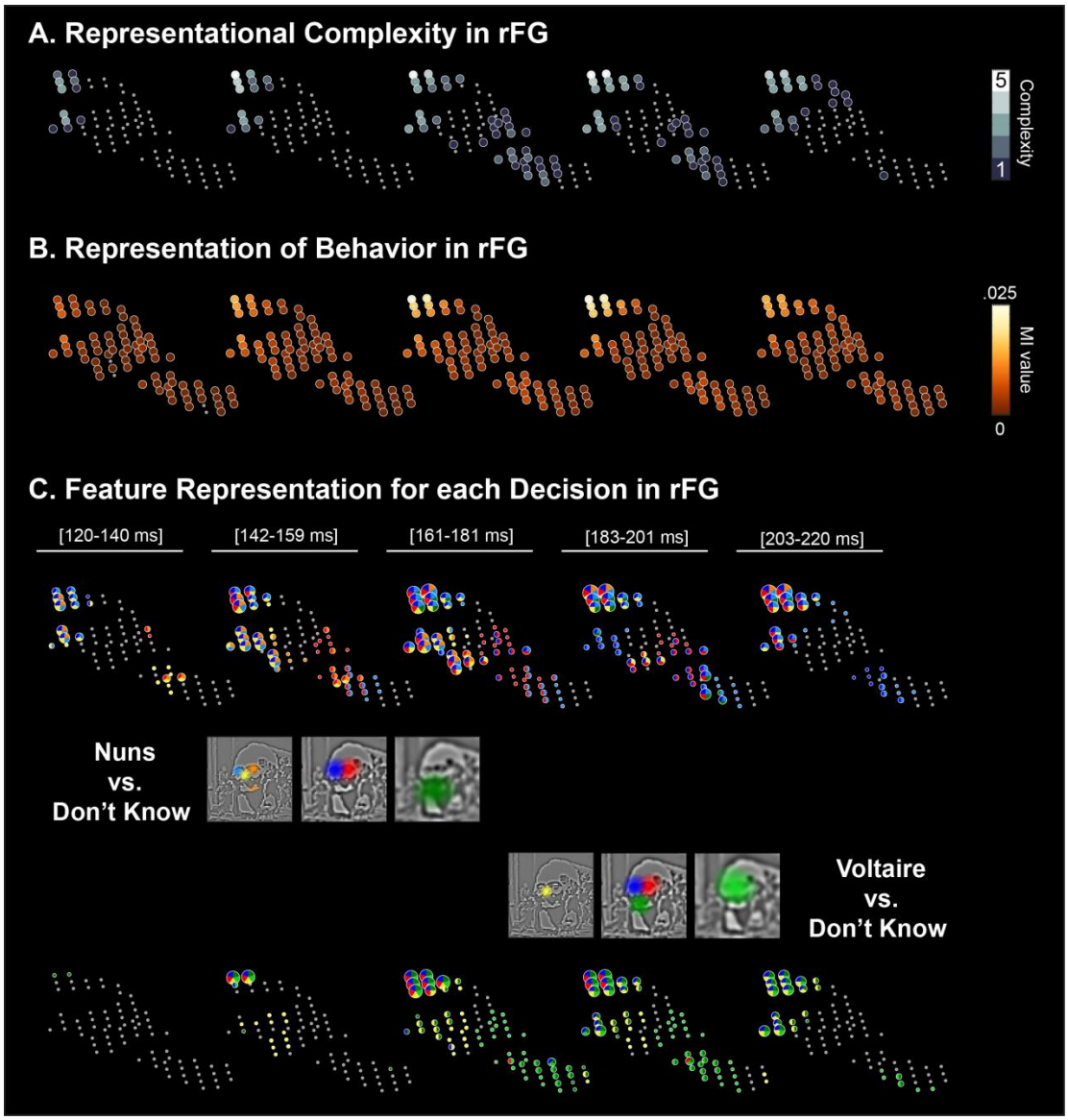

Figure S 4-10 Dynamic Construction of Representations for Behavior in rFG (Participant 2). Same caption as in Figure S4-9.

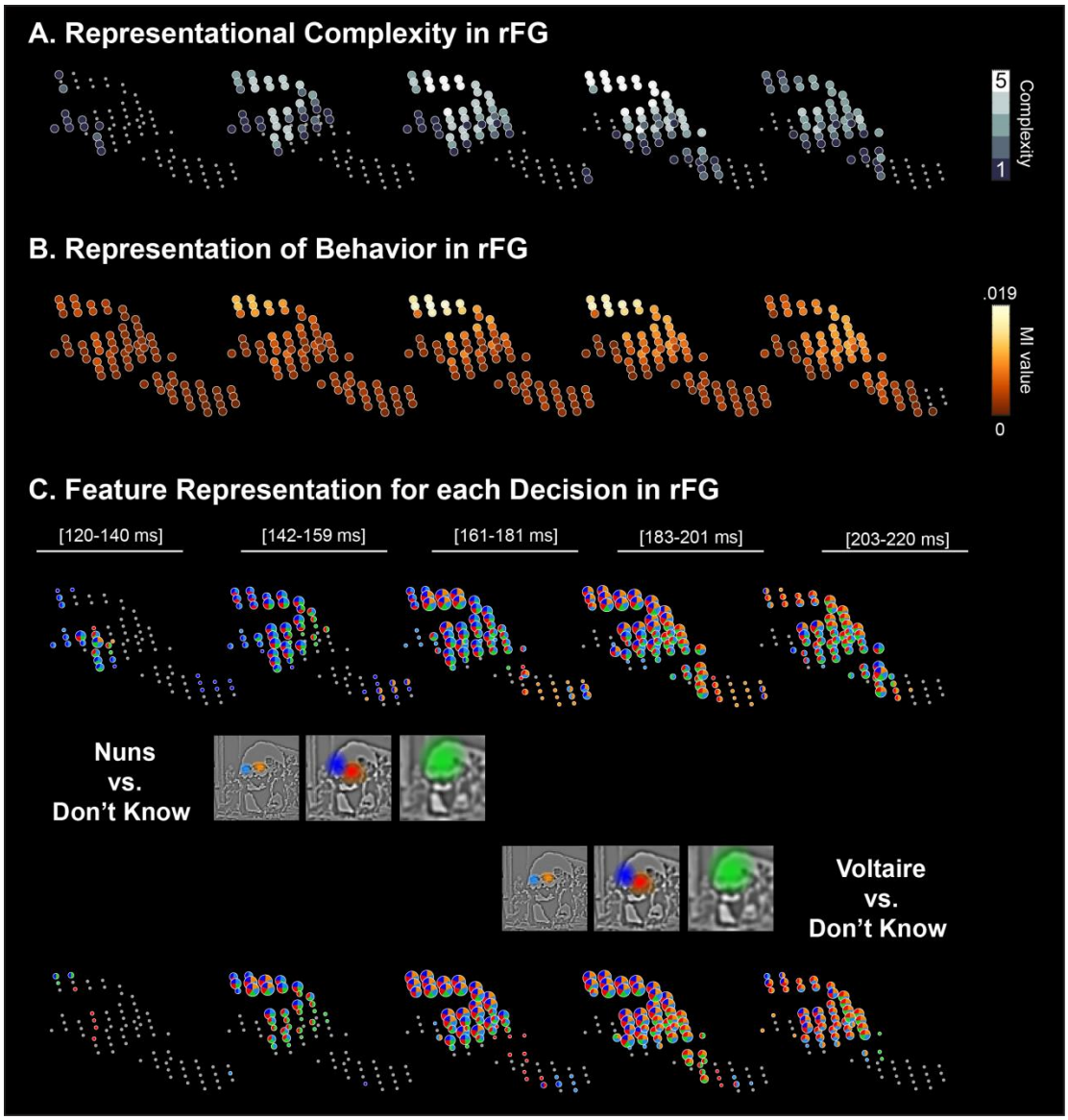

Figure S 4-11 Dynamic Construction of Representations for Behavior in rFG (Participant 3). Same caption as in Figure S4-9.

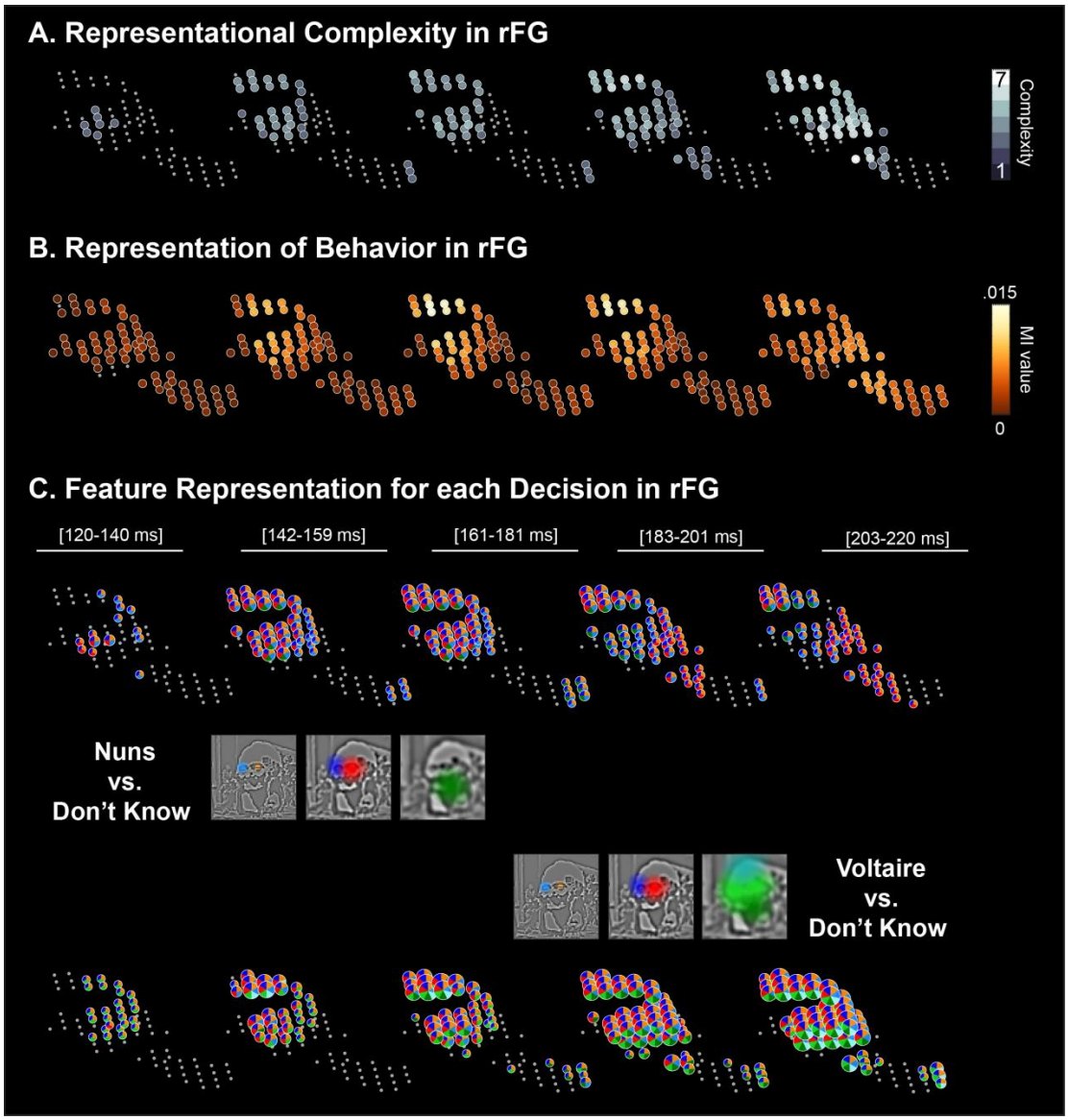

Figure S 4-12 Dynamic Construction of Representations for Behavior in rFG (Participant 4). Same caption as in Figure S4-9.

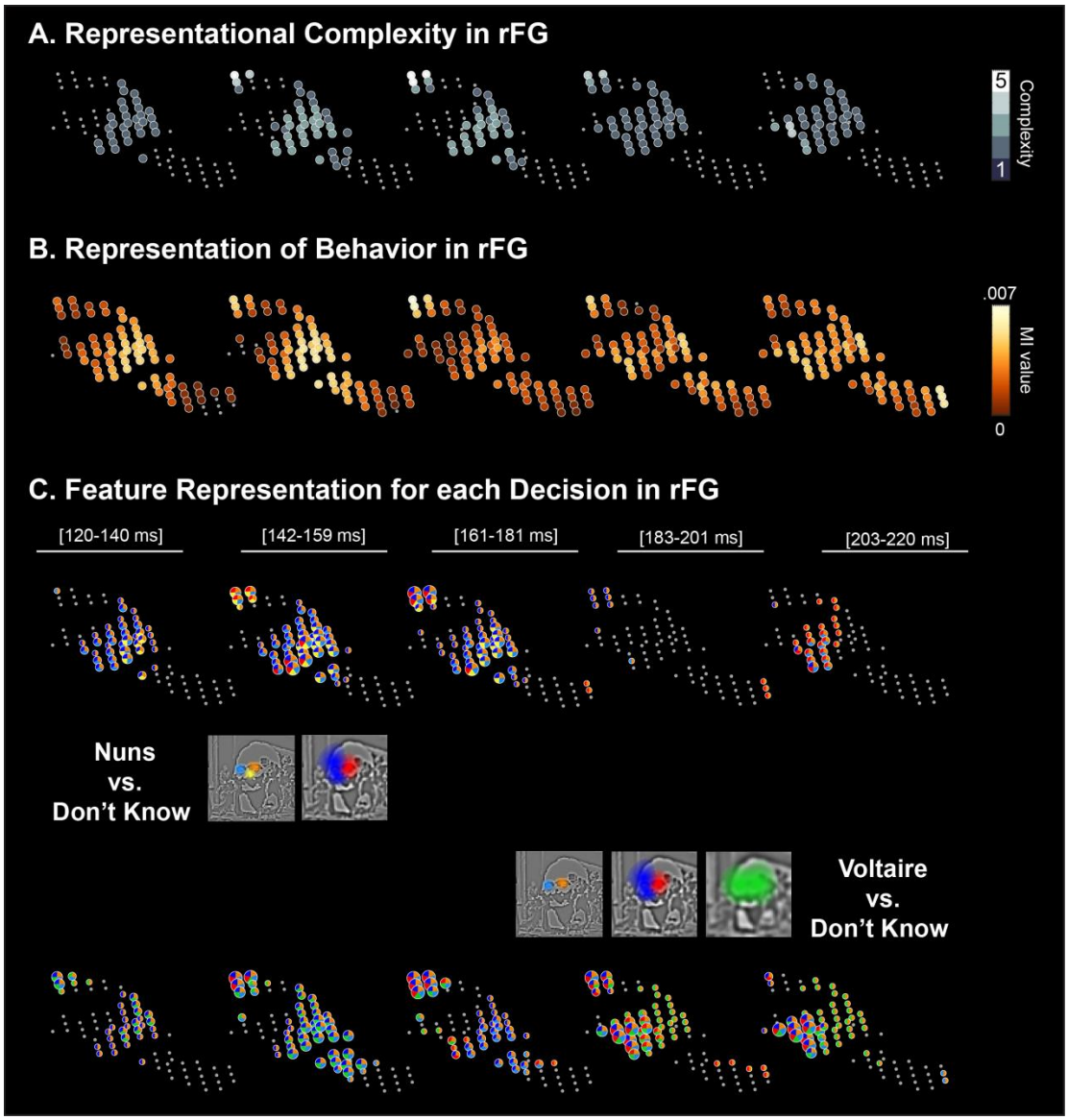

Figure S 4-13 Dynamic Construction of Representations for Behavior in rFG (Participant 5). Same caption as in Figure S4-9.

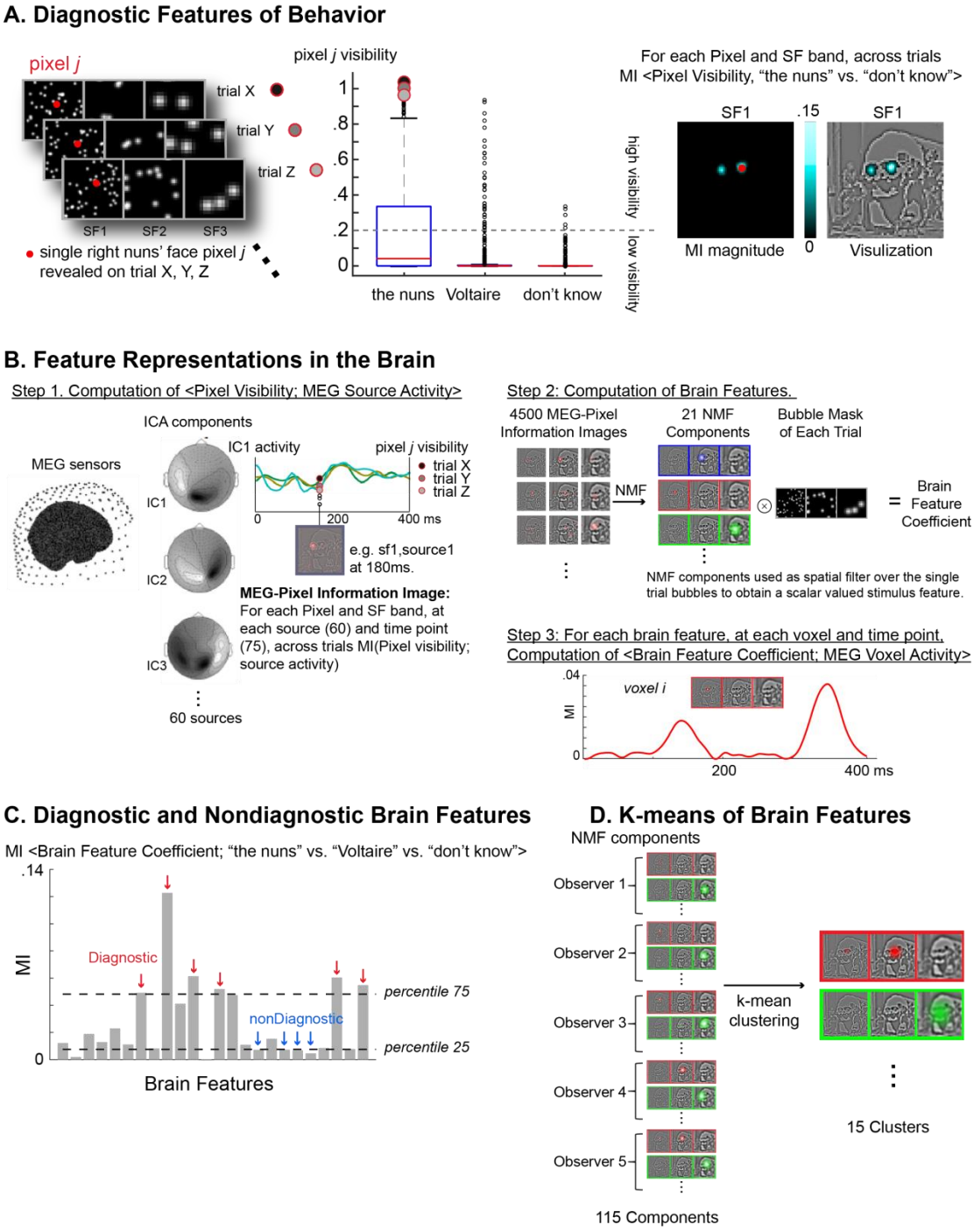

Figure S 4-14 Schematic of Analysis Pipeline. A. Diagnostic Features of Behavior illustrates the single trial MI calculation of the relationship <Information Samples; "the nuns" vs. "don't know">. B. Feature Representations in the Brain illustrates the computations leading to Brain Features (i.e. steps 1 & 2) and the calculation of MI <Brain Features; MEG Voxel Activity> (step 3). C. Diagnostic and Nondiagnostic Brain Features illustrates the determination of nondiagnostic and diagnostic brain features from the distribution of MI <Brain Feature Coefficient; "the nuns" vs. "Voltaire" vs. "don't know">. D. K-means of Brain Features illustrates k-means clustering of brain features across observers to display group level representations of similar features on MEG voxels.

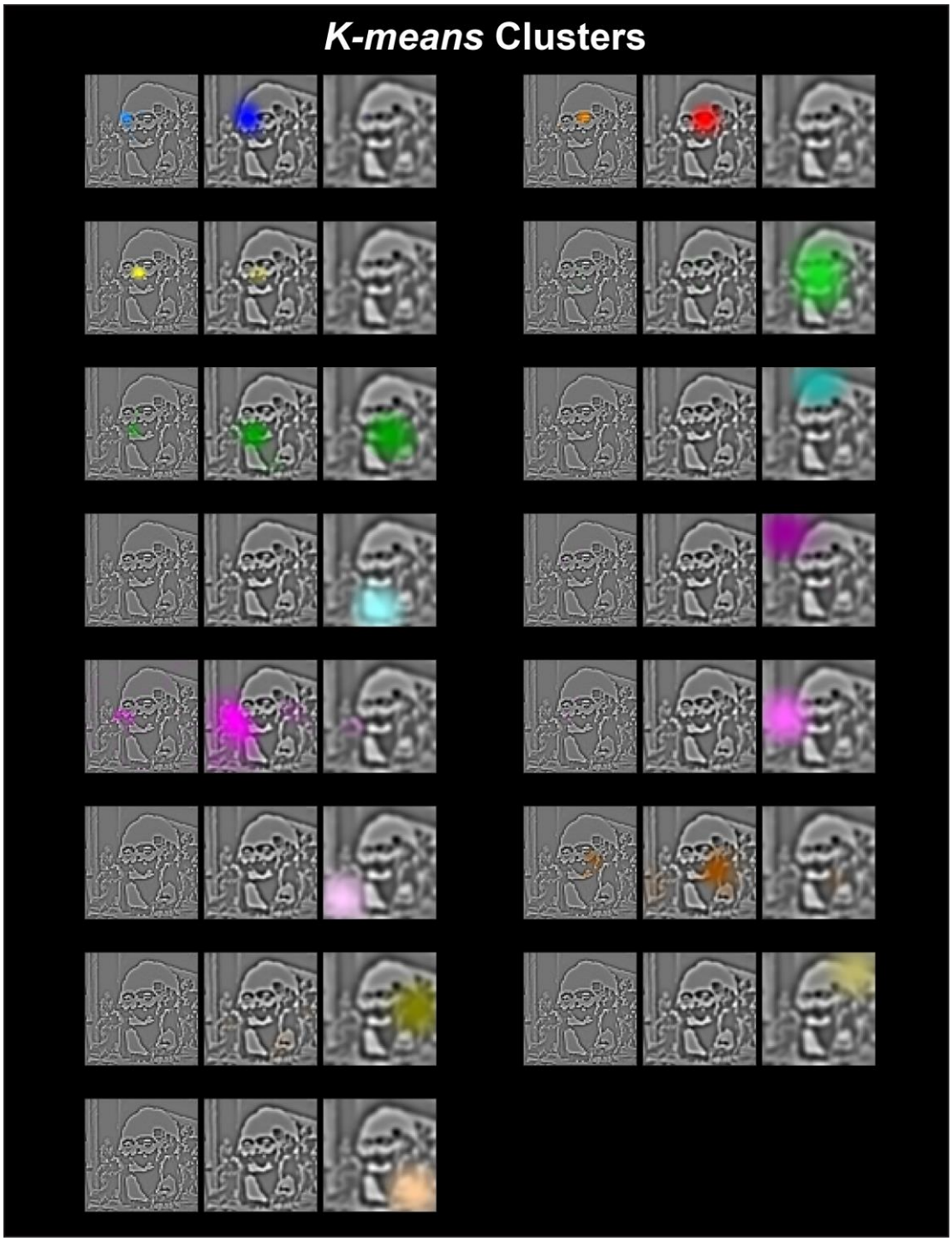

Figure S 4-15 K-means Analysis. Color-coded, clustered brain features (i.e. k-means) computed from the brain features of 5 observers.

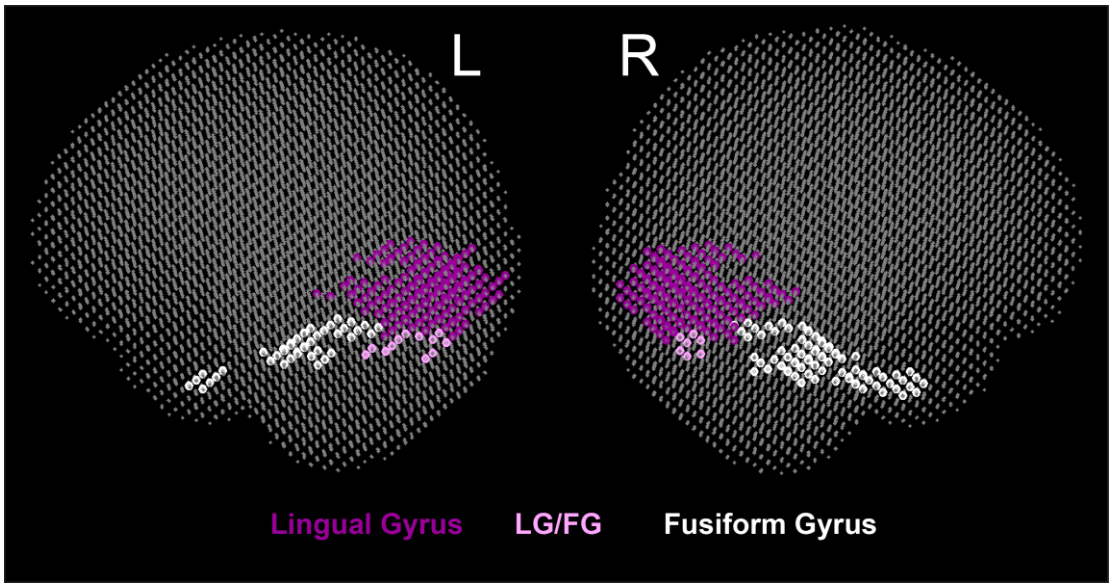

Figure S 4-16 Location of LG/FG Voxels. The dark purple scatters show lingual gyrus (LG) voxels; the light purple scatters show LG/FG voxels which are fusiform gyrus voxels located next to lingual gyrus voxels; the white scatters show the well-demarcated FG voxels that we included in our analysis of feature representations for behavior.

5 General Discussion

Mental representations stored in memory are critical to guide the information processing mechanisms of cognition. In this thesis, I concretize the information contents of mental representations by investigating the diagnostic information for different visual categorization tasks.

In **Study 1**, I modelled the diagnostic information of familiar face identification. Using a novel 3D face information generator, I reverse correlated the representational contents of 4 familiar faces in 14 individual participants. Further analyses reveal that the diagnostic contents across these participants converge on a subset of faithful features, which maximally distinguish each identity from their categorical averages (i.e., the average face with the same age/gender/ethnicity as the identity) and enable a new group of validators to generalize the identification to different views, age and even the kinship task.

In **Study 2**, I modelled the diagnostic information for facial expressions of emotion recognition. I used the models that characterize the mental representations of six facial expressions of emotion in individual participants. I selected a subset as the validated models based on their categorization accuracy in a new group of validators. A cluster analysis of these validated models derived main variants for each emotion and their probability to be produced. Using a Bayesian classifier, I demonstrated that the diagnostic contents of facial expressions of emotion comprise these main variants and their probability of occurrence.

In **Study 3**, I tracked how our brain dynamically reduced the nondiagnostic information and constructed the diagnostic information for a scene categorization task. With a new information theoretic framework -- Contentful Brain and Behavior Imaging -- I first teased apart two types of stimulus information represented in the brain: that which supports decision behavior (i.e., diagnostic) and that which does not (nondiagnostic). Then I tracked the dynamic representations of both in the brain, using the magneto-encephalography (MEG) that is advantageous to both temporal and spatial resolution. My results reveal a many to fewer dimensions reduction of stimulus representation along the occipital-ventral pathway: the nondiagnostic features are rapidly collapsed at the occipito-ventral junction around 170ms, whereas the diagnostic features are transmitted progressively into right

fusiform gyrus where multiple diagnostic features are integrated for distinct behavior between 161 and 201 ms, post-stimulus.

Together, using elegant information sampling techniques and the reverse correlation method, my studies identified the diagnostic information for three different visual categorization tasks: familiar face identification, facial expressions of emotion recognition, scene categorization. This task-specific diagnostic information formalizes observers' mental representation of different visual categories (i.e., familiar faces, six emotions, and objects in the scene), which provides the critical link to derive a complete information processing mechanism for visual categorizations. The CBBI framework further enables the information processing interpretation in neural (neutrally-inspired) activity. In this section, I will discuss some points regarding the psychological status of diagnostic information (see ***5.1***), information processing implemented in the brain (see ***5.2***), and practical concerns about using the data-driven approach to uncover the diagnostic information (see ***5.3*** and ***5.4***).

5.1 Diagnostic Information: the Experienced-based Representation of the Physical World

How we represent the physical world is strongly embedded in the way we interact with it (*Barsalou, 2009*). In visual task, individuals' experience shapes their inner representation about the input stimuli. This fact is massively documented by a wide variety of studies, involving perceptual learning (*Ahissar & Hochstein, 1997, 2004; Dosher & Lu, 1998; Fine & Jacobs, 2002; J. Gold, Bennett, & Sekuler, 1999*) and concept learning (*Goldstone, 1994; Livingston, Andrews, & Harnad, 1998; Schyns & Rodet, 1997; Soto & Ashby, 2015; Tanaka, Curran, & Sheinberg, 2005*) on normal people and the investigations on the people who have been deprived of experience from early life (*Bouvrie & Sinha, 2007*). The key idea is that during a course of training, observers learn distinctive features by using the information that improves their behavioral output (e.g., detection and categorization performance) efficiently, and as a result the stimulus representation redundancy is reduced. Such acquired feature distinctiveness is memorized by observers, which is not fixed but dynamically updated to adjust to new contingencies. Culture-specific mental representation *(Blais et al., 2008; Chua, Boland, & Nisbett, 2005; Jack, Caldara, et al., 2012; Jack, Garrod, et al., 2012; Kelly et al., 2007; Nisbett & Miyamoto, 2005*) provides a good illustration of experience-based representation

in real life, and the development of perceptual expertise (*Palmeri, Wong, & Gauthier, 2004; Tanaka & Taylor, 1991*) and the face familiarity effect (*Ellis, Shepherd, & Davies, 1979*) in visual cognition are two practical cases illustrate the representation plasticity outside the laboratory scenario.

Taken together, individuals' knowledge of the external event is dynamically developed when they interact with the environment. This informs the observers to update their weights on features selection and forms the effective representation to deal with upcoming events. As a result, the diagnostic information is formalized as only a reduced set of the full information input and constructs observers' dynamic psychological space of the physical world. In this stance, a precise modelling about the emergence/development of psychological space for visual categorization is necessary, which enable us to better theorize the psychological status of these dimensions. The methodology framework of my thesis provides a practical way to deal with this issue. For example, we can use advanced generative models to tightly control the physical stimuli, reverse correlated the diagnostic information underlying a task demand, and track the formation of psychological dimensions (substantialized as diagnostic information) through training. With the CBBI framework (introduced in Chapter 3), we can further clarify the information processing development from neural processing. This approach is generalizable to face, object and scene categories in visual recognition, as well as other categories in different perceptual mode (e.g. auditory, olfactory and tactile).

5.2 The Computable Mind for Visual Categorization

An influential idea links cognition, the brain, and the theory of computation is that we explain human behavior in terms of information processing mechanisms. Such information emphasis makes it possible to track and compute the transformation between stimuli input and behavioral output, and formalize the cognitive processing as a series of computable inner-states implemented in the functional architectures of the human brain. To offer an analysis of such information processing systems and the underlying neural mechanisms, Marr provided his three levels of analysis: the computational, the algorithmic, and the implementational (*Marr, 1982*). The computational level defines the most abstract level, which describes the overall goal of the computation (e.g., identify a familiar face), what is computed (e.g., visual information) and the logical to achieve the computation (e.g., using a subset of face information among all visual information).

The algorithmic level decomposes the abstract goal into specific procedures, by defining the representation of input (e.g., 3D vertices position for face shape and 2D RGB pixels for face texture), and output (e.g., an identity name), and the algorithms used to transform the former into the later. The implementational level deals with the "the details of how the algorithm and representation are realized physically" (p. 25 in *Marr, 1982*), for example in the brain.

5.2.1 Computational goal

The first *computational* level plays a crucial role in the analysis because it defines the computation goal, i.e., what is computed from the stimuli input. The models of diagnostic information should be constructed as the abstract information goals that the visual system predicts under different task demands, and mapped onto this computation level. I term the diagnostic information as the 'abstract information goal' because it has to be broken down into global and local constitutes according to the constraints of representation and implementation at each level of the visual hierarchy. For example, we would hypothesize that the diagnostic identity components in Study 2 (cf. Figure, 2-8A) are broken down, bottom to top, into the representational language of V1—i.e., as representation in multi-scale, multi-orientation Gabor-like, retinotopically mapped receptive fields (*Kay et al., 2008; F. W. Smith & Muckli, 2010*); at intermediate levels of processing, as the sort of local surface patches (*Kubilius et al., 2014; Peirce, 2015*) that we reveal, and at the top level as the combinations of surface patches that enable identification and resemblance responses. Under a framework of top-down prediction (*A. Clark, 2013; K. J. Friston & Kiebel, 2009*), the abstract information goal of a familiar face identity should trim, in a top-down manner, the fully-mapped but redundant information on the retina into the task-relevant features that are transferred along the occipital to ventral/dorsal visual hierarchy. The study reporting in Chapter 4 provides the first step to characterize the spatiotemporal dynamics of such information reduction from neural processing.

5.2.2 Theoretic Algorithm

Decision-theoretic models provide a good framework to algorithmize the visual categorization, because they do not only propose a series of inner states underlying the decision (i.e. evidence representation, evidence accumulation and integration, and decision making) but also specify their corresponding operations

in computational level and the implementation in neural level (see a general review in 1.3.3.1 and detailed discussions in *J. I. Gold & Shadlen, 2007; Philiastides & Heekeren, 2009*). On such model basis, my work contributes further by precisely quantifying the represented evidence that has to be task-targeted, which is not explicitly specified in theoretic models for perceptual decision. In this section, I will discuss a possible operation that happens earlier and enables the task-relevant evidence representation based on selective attention and diagnostic recognition.

Visual systems rely on selective attention to filter out unwanted information (by inhibition) – i.e. the nodiagnostic information, and focus on the information relevant to current goals (by amplification) – i.e. the diagnostic information (see reviews *Hillyard et al., 1998*), to facilitate the perception (*Desimone & Duncan, 1995*). The Guided-Search model of selective attention by Wolfe (*1994*) provides a good computational account to explain the information reduction under the diagnostic recognition. This model introduces the idea of a top-down feature map which in its essence weights the sensory inputs according to their relevance or diagnosticity to the task goal. This model theorizes a two-stage processing of visual information selection (see also *Treisman & Gelade, 1980*), which can formalize the reduction of nondiagnostic information found in my results. The first stage is pre-attentive and encodes in a spatially parallel way the presence and physical salience of simple visual features mapped on the retina. In this stage, the visual system creates a retinotopic map reflects the bottom-up activation value of each feature, which can explain the earlier representation of both diagnostic and nondiagnostic features. Then the bottom-up feature values are combined with the top-down feature values according to their task-relevance to generate the "saliency feature map". In the second and attentive stage, the visual system searches information in a spatially serial way according to the feature values defined by the 'saliency map'. The greater the activation at a location, the more likely the attention will be directed to that location earlier. The serial search terminates if the features are accumulated enough for a successful recognition, which defines the diagnostic features I reconstructed from the reverse correlation task.

It is interesting that we can notice the similarity between the Guide Search model of selective attention and the computation models of perceptual decision. That is, both two specify the process of information accumulation before the

decision forms, i.e. a diffusion model that visual information is continuously accumulated until one of N response criteria (e.g. in N-AFC task) is reached. What the selective attention model informs the decision model are: 1) it specifies the starting point of the evidence accumulation (i.e. the second stage) and 2) it increases the drift rate (i.e. achieve to the response criteria in a short decision time) due to the accumulation of task-relevant features.

5.2.3 Neural Implementations

Based on the theoretic algorithm discussed in 5.2.2, I now would like to discuss the nondiagnostic information reduction using a sensory gain-control mechanism underlying selective attention (*Hillyard et al., 1998*). The sensory 'gain control' enhances the excitability of extrastriate neurons coding attended features (i.e. diagnostic features in my case) and suppress the excitability (or reduce the gain) of those coding ignored features (i.e. nondiagnostic features). In EEG research, three main components are usually found in the selective attention tasks that compare the attended vs. unattended stimuli (*Hillyard & Anllo-Vento, 1998*). These components are termed as C1, P1, and N1 and have been shown to play distinct roles in sensory gain control.

C1 is the earliest ERP component with an onset latency of 50-60ms, but it does not show any significant changes between attended and unattended stimuli (*V. P. Clark & Hillyard, 1996; Gomez Gonzalez, Clark, Fan, Luck, & Hillyard, 1994; Johannes, Munte, Heinze, & Mangun, 1995; Wijers, Lange, Mulder, & Mulder, 1997*). Source localization of C1 points its neural generator in occipital regions *(V. P. Clark & Hillyard, 1996; Gomez Gonzalez et al., 1994*), and C1 systematically changes its topography and as a function of stimuli position in a manner of retinotopic organization (i.e. contra-lateral mapping, *V. P. Clark, Fan, & Hillyard, 1995*). These findings are consistent with what I found for the early contra-lateral representation (i.e. < 170ms) of both diagnostic and nondiagnostic features in occipital regions (c.f. Figure 4-1C and Figure 4-2).

The P1 component starting from 80ms post-stimuli appears to be the earliest neural attention effects, i.e. larger amplitude for attended stimuli than unattended ones (*V. P. Clark & Hillyard, 1996; Johannes et al., 1995; Rugg, Milner, Lines, & Phalp, 1987; Valdes-Sosa, Bobes, Rodriguez, & Pinilla, 1998; Wijers et al., 1997*). This component has been demonstrated to reflect a

suppression of input processing at ignored (task-irrelevant) locations to reduce the interference and processing cost (*Luck & Hillyard, 1995*). The source localization techniques point the generator of P1 primarily in lateral extrastriate cortex near the border of Brodmann's Areas 18 and 19 (*V. P. Clark et al., 1995*), which together with its time window (80 – 160ms) are consistent with where and when I found the reduction of nondiagnostic features in my MEG study. In a recent study, Zanto et al. (*2011)* used the repetitive transcranial magnetic stimulation (rTMS) to perturb the function of inferior frontal junction prior to participants performing a selective-attention task, and found such perturbation reduced the suppression degree of ignored stimuli and diminishes P1 amplitude difference between the processing of attended vs. ignored stimuli to nonsignificant. This study provided direct evidence for the top-down inhibition from prefrontal cortex to early visual regions, indicating a neural network underlying the information reduction (of non-diagnostic features). In my MEG study, I indeed found the prefrontal activation in one observer, which peaks at the time when nondiagnostic features are reduced below threshold (c.f. Figure 4-3). This finding implies a top-down modulation from frontal regions to occipital-temporal junction to inhibit the representation of nondiagnostic information during visual categorization. However, the prefrontal activation is missing in other four observers, which might be due to the very conservative threshold with multiple corrections over 12,773 voxels and more than 20 features.

The third component N1 also shows larger amplitude for attended vs. ignored stimuli and appears in the 120 – 180ms post-stimuli range *(V. P. Clark & Hillyard, 1996; Johannes et al., 1995; Rugg et al., 1987; Valdes-Sosa et al., 1998; Wijers et al., 1997*). This time window is roughly the same as my study that showed feature integration in right fusiform gyrus. Different from P1, N1 attention effect reflects the amplification of input processing at the attended location and therefore facilitates the discriminative processing *(Luck, 1995; Luck & Hillyard, 1995*). Thus, this component should relate to the representation and integration of task-relevant evidence (i.e. diagnostic features) for the perceptual decision. Source localization suggests multiple generators of the N1 effect, including posterior temporal, parietal and frontal regions; and the frontal and parietal ones with latency between 130 and 140ms appear earlier than the posterior temporal one with latency between 165 – 175ms (V. P. Clark et al., 1995). The posterior temporal one should relate to the N170, the Event Related Potential ~170 ms post-stimulus commonly associated with visual categorizations (*Bentin et al., 1996;*

Cichy et al., 2014). Together with my results showing the increased number of diagnostic features represented in right fusiform gyrus and other studies showing increased activity in frontal and parietal regions as more evidence accumulated for decision, such spatial-temporal profile of N1 families should indicate a top-down modulation from prefrontal and parietal regions for the integration of diagnostic features for visual categorization in ventral temporal cortex *(Heekeren et al., 2004; Ploran et al., 2007*).

5.3 Reverse Correlated Diagnostic Information: Memory Representations or Task Representations?

A general question with reverse correlation tasks is whether the resulting models represent the diagnostic information of a particular visual category or the task from which the model was reconstructed (see discussion in *Schyns, 1998*). Here, let me use the face identity experiment (Chapter 2) to illustrate.

Since the reverse correlation method examines face recognition under highly constrained and controlled conditions, there are concerns about the ecological validity of the technique. First, in this study, recognition of four familiar persons was tested with faces shown in a single frontal pose. Hence, the findings could be idiosyncratic to the face familiarity in one image rather than the robust recognition of a familiar person across many images that the face recognition works in our everyday lives. Second, to derive the mental representation model, the method required 90 blocks of 20 classification trials per identity (7200 total trials). This experimental procedure could change the face recognition process itself, e.g., participants adopt strategies driven by the task demands of the experiment that they would not normally apply in real-world face recognition.

To justify the reverse correlation approach, I run validation experiments afterward, which contributed to such mental representation vs. task representation debate by showing that the identity information reconstructed in one task had efficacy in other tasks that involved identity. Importantly, the tasks were designed to test two classes of factors: ambient and categorical. For example, I showed that the identity component extracted in one ambient viewpoint (full face, 0 deg) could be used to generalize identification of the same face under two new ambient viewpoints (-30 and +30 deg of rotation in depth). I also showed that the identity component extracted for identities (all < 40 years of age) generalized to older age

(80 years). Furthermore, I showed that though extracted from a given sex, the identity component would generalize to another sex, a kinship task. Hence, there is no dramatic difference due to the effect of task of extraction of the identity component. Rather, the extracted representational basis is useful for all tasks tested, whether using ambient or categorical factors of face variance. This therefore suggests that using the reverse correlation approach I have tapped into some essential information about familiar face identity. However, I acknowledge that the generalizations I demonstrated might still be a function of an interaction between the nature of memory and the similarity task from which I estimated the identity component. The component could have differed had the task been more visual than memory based (e.g., identification of the same face under different orientations, or a visual matching task) and my experiments might not have derived an identity component that enabled such effective generalization. To achieve the robust estimation about the mental representations, future work should also take more task flexibility into consideration.

5.4 Reverse Correlation: What Information Should We Sample?

What information we should sample relies on at what level we aim to understand the diagnostic information. If we aim to know which part of the information in a 2D image drives a perceptual decision (cf. Dali ambiguous painting in Chapter 4), we can sample the information visibility to observers. On this basis, Gosselin and Schyns (*2001*) designed the Bubbles technique, which samples the pixel visibility at different spatial frequency using contiguous bubbles. Bubbles sampled spatial frequency because spatial filtering is an early stage of visual processing (*Ginsburg, 1978*), which enables the representation of a wide range of visual information, ranging from the fine details to large coarse parts. Another practice to sample the visibility is adding noise to stimuli, using either the white noise (random greyscale pixel values, e.g. *Jack, Caldara, et al., 2012*) or the structured (Gabor) noise (van Rijsbergen, Jaworska, Rousselet, & Schyns, 2014). In either way (i.e., Bubbles or adding noise), few assumptions are set about the structure of the stimuli population (cf. the multidimensional features space), thus we can call it the ‘bottom-up’ sampling.

However, image sampling via pixel visibility is relevantly a brute force approach. Under this approach, we understand the diagnostic information in a

manner of presence vs. absence, but we cannot know how the revealed diagnostic features relate the structure of the visual events, e.g., how a diagnostic face features distribute in an information space describing the variance all white Caucasian faces. If we aim to investigate the diagnostic information at the structure level, we need to sample the generative basis that produces the 2D images (cf. Generative Model of Face Identity in Chapter 2 and GFG in Chapter 3). Unlike a bottom-up sampling, generative model is designed rather in a top-down manner. It is created fully based on the explicit hypothesis of the visual information which we researchers believe to support the categorization and is also sufficiently realistic to engage the visual system. For example, we model the 3D face shape and colored texture as the generative basis of individual face identification (Chapter 2), and model biologically plausible facial movement as the generative basis of facial expressions (Chapter 3). Bear in mind, sampling information by a generative model will limit the investigation to the information dimensions the model can generate.

6 Concluding Remarks

I started with a perspective that we can model the visual categorization as information processing flows implemented in the brain, with the information contents specified along with the stimuli input to behavioral output transition. I discussed the mental representations as the critical component to derive the complete information processing explanations of visual categorizations, and discussed the task-specific diagnostic information as a precise estimation of mental representations of different categories, which together motived me to carry out three studies. In my first and second behavioral studies, I modelled the diagnostic information that guides the visual information processing for familiar face identification and facial expressions of emotion recognition. In my third study, I tracked the diagnostic as well as the nondiagnostic information represented in the brain, and documented the input-to-output transition as many-to-fewer dimensions of information reduction along the occipito-ventral pathway. The approach and results I provided open new research avenues for the interplay between visual information, categorization tasks and their implementation as information processing mechanisms in the brain.

List of References

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: Evidence and implications. *Neuron, 21*(2), 373-383. doi:Doi-10.1016/S0896-6273(00)80546-2

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387*(6631), 401-406. doi:DOI 10.1038/387401a0

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci, 8*(10), 457-464. doi:10.1016/j.tics.2004.08.011

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage, 23*(3), 905-913.doi:10.1016/j.neuroimage.2004.07.060

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *J Exp Psychol Learn Mem Cogn, 14*(1), 33-53.

Ashby, F. G., & Maddox, W. T. (1992). Complex Decision Rules in Categorization - Contrasting Novice and Experienced Performance. *Journal of Experimental Psychology-Human Perception and Performance, 18*(1), 50-71. doi:Doi 10.1037//0096-1523.18.1.50

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends Cogn Sci, 11*(7), 280-289. doi:10.1016/j.tics.2007.05.005

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., . . . Halgren, E. (2006). Top-down facilitation of visual recognition. *Proc Natl Acad Sci U S A, 103*(2), 449-454. doi:10.1073/pnas.0507062103

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos Trans R Soc Lond B Biol Sci, 364*(1521), 1281-1289. doi:10.1098/rstb.2008.0319

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*(6), 551-565. doi:DOI-10.1162/jocn.1996.8.6.551

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture Shapes How We Look at Faces. *PLoS One, 3*(8). doi:ARTN e302210.1371/journal.pone.0003022

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Siggraph 99 Conference Proceedings*, 187-194.

Bouvrie, J. V., & Sinha, P. (2007). Visual object concept discovery: Observations in congenitally blind children, and a computational approach. *Neurocomputing, 70*(13-15), 2218-2233. doi:10.1016/j.neucom.2006.01.035

Bullmore, E. T., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10*(3), 186-198.doi:10.1038/nrn2575

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments Against a Configural Processing Account of Familiar Face Recognition. *Perspect Psychol Sci, 10*(4), 482-496. doi:10.1177/1745691615583129

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*(8), 641-651. doi:10.1038/nrn1724

Carey, S., & Diamond, R. (1977). From Piecemeal to Configurational Representation of Faces. *Science, 195*(4275), 312-314. doi:DOI 10.1126/science.831281

Chang, C. H., Nemrodov, D., Lee, A. C. H., & Nestor, A. (2017). Memory and Perception-based Facial Image Reconstruction. *Scientific Reports, 7*. doi:ARTN 649910.1038/s41598-017-06585-2

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell, 169*(6), 1013-1028 e1014. doi:10.1016/j.cell.2017.05.011

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proc Natl Acad Sci U S A, 102*(35), 12629-12633.doi:10.1073/pnas.0506162102

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports, 6*. doi:ARTN 2775510.1038/srep27755

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience, 17*(3), 455-462. doi:10.1038/nn.3635

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. doi:10.1017/S0140525x12000477

Clark, V. P., Fan, S., & Hillyard, S. A. (1995). Identification of early visual evoked potential generators by retinotopic and topographic analyses. *Human Brain Mapping, 2*, 170 - 187.

Clark, V. P., & Hillyard, S. A. (1996). Spatial selective attention affects early extrastriate but not striate components of the visual evoked potential. *J Cogn Neurosci, 8*(5), 387-402.doi:10.-1162/jocn.1996.8.5.387

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 681-685. doi:Doi 10.1109/34.927467

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Cukur, T., Huth, A. G., Nishimoto, S., & Gallant, J. L. (2013). Functional Subdomains within Human FFA. *Journal of Neuroscience, 33*(42), 16748-16766. doi:10.1523/Jneurosci.1259-13.2013

Darwin, C. (1999). *The expression of the emotions in man and animals (3rd ed.)*. London: Fontana.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual-Attention. *Annual Review of Neuroscience, 18*, 193-222. doi:DOI 10.1146/annurev.neuro.18.1.193

Dosher, B. A., & Lu, Z. L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences of the United States of America, 95*(23), 13988-13993. doi:DOI10.1073/pnas.95.2-3.13988

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology, 92*, 53-78. doi:Doi 10.1348/000712601162103

Eger, E., Schweinberger, S. R., Dolan, R. J., & Henson, R. N. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *Neuroimage, 26*(4), 1128-1139. doi:10.1016/j.neuroimage.2005.03.010

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation, 19*, 207-283.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: Investigator's Guide*: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., Osullivan, M., Chan, A., Diacoyannitarlatzis, I., Heider, K., . . . Tzavaras, A. (1987). Universals and Cultural-Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology, 53*(4), 712-717. doi:Doi10.1037/0-022-3514.53.4.712

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science, 164*(3875), 86-&. doi:DOI 10.1126/science.164.3875.86

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin, 128*(2), 203-235.

Elfenbein, H. A., Beaupre, M., Levesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion, 7*(1), 131-146. doi:10.1037/1528-3542.7.1.131

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of Familiar and Unfamiliar Faces from Internal and External Features - Some Implications for Theories of Face Recognition. *Perception, 8*(4), 431-439. doi:DOI 10.1068/p080431

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron, 23*(1), 115-125. doi:Doi 10.1016/S0896-6273(00)80758-8

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601. doi:Doi 10.1038/33402

Erens, R. G., Kappers, A. M., & Koenderink, J. J. (1993). Perception of local shape from shading. *Percept Psychophys, 54*(2), 145-156.

Estes, W. K. (1986). Array Models for Category Learning. *Cognitive Psychology, 18*(4), 500-549. doi:Doi 10.1016/0010-0285(86)90008-3

Fairhall, S. L., & Ishai, A. (2007). Effective connectivity within the distributed cortical network for face perception. *Cerebral Cortex, 17*(10), 2400-2406. doi:10.1093/cercor/bhl148

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning across tasks: A review. *Journal of Vision, 2*(2), 190-203. doi:Artn 510.1167/2.2.5

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195-1201. doi:10.1038/nn.2889

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974-981. doi:10.1038/nn.3402

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), 159-170. doi:DOI10.1068/p30-12

Friston, K. (2008). Hierarchical models in the brain. *Plos Computational Biology, 4*(11), e1000211. doi:10.1371/journal.pcbi.1000211

Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1521), 1211-1221.doi:10.1098-/rstb.2008.0300

Fukushima, K. (1988). Neocognitron - a Hierarchical Neural Network Capable of Visual-Pattern Recognition. *Neural Networks, 1*(2), 119-130. doi:Doi 10.1016/0893-6080(88)90014-7

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568-573. doi:Doi 10.1038/9224

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience, 12*(3), 495-504. doi:Doi 10.1162/089892900562165

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), 574-580. doi:Doi 10.1038/9230

Ginsburg, A. P. (1978). *Visual Information Processing Based on Spatial Filters Constrained by Biological Data.* (Ph.D. dissertation), University of Cambridge, Cambridge.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature, 402*, 176-178.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574. doi:10.1146/annurev.neuro.29.051605.113038

Goldstone, R. (1994). Influences of Categorization on Perceptual Discrimination. *Journal of Experimental Psychology-General, 123*(2), 178-200. doi:Doi 10.1037//0096-3445.123.2.178

Gomez Gonzalez, C. M., Clark, V. P., Fan, S., Luck, S. J., & Hillyard, S. A. (1994). Sources of attention-sensitive visual event-related potentials. *Brain Topography, 7*(1), 41-51.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research, 41*(17), 2261-2271. doi:Doi 10.1016/S00426989(01)0-0097-9

Gosselin, F., & Schyns, P. G. (2002). RAP: a new framework for visual categorization. *Trends Cogn Sci, 6*(2), 70-77.

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology, 13*(2), 159-166. doi:10.1016/S0959-4388(03)00040-0

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience, 7*(5), 555-562. doi:10.1038/nn1224

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10-11), 1409-1422. doi:Doi 10.1016/S0042-6989(01)00073-6

Grill-Spector, K., Sayres, R., & Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature Neuroscience, 9*(9), 1177-1185. doi:10.1038/nn1745

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience, 15*(8), 536-548. doi:10.1038/nrn3747

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., . . . Schoffelen, J. M. (2013). Good practice for conducting and reporting MEG research. *Neuroimage, 65*, 349-363. doi:10.1016/j.neuroimage.2012.10.001

Guclu, U., & van Gerven, M. A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci, 35*(27), 10005-10014. doi:10.1523/JNEUROSCI.5023-14.2015

Hall, E. (1966). *The Hidden Dimension* Garden City, NY: Doubleday.

Hanson, S. J., & Schmidt, A. (2011). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. *Neuroimage, 54*(2), 1715-1734. doi:10.1016/j.neuroimage.2010.08.028

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proc Natl Acad Sci U S A, 111*(10), E962-971. doi:10.1073/pnas.1312567111

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233. doi:Doi 10.1016/S1364-6613(00)01482-0

Heekeren, H. R., Marrett, S., Bandettini, P. A., & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature, 431*(7010), 859-862. doi:10.1038/nature02966

Hernandez, A., Zainos, A., & Romo, R. (2002). Temporal evolution of a decision-making process in medial premotor cortex. *Neuron, 33*(6), 959-972. doi:Doi 10.1016/S0896-6273(02)00613-X

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior, 24*(4), 265-283. doi:Doi 10.1023/A:1006623213355

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America, 95*(3), 781-787. doi:DOI 10.1073/pnas.95.3.781

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 353*(1373), 1257-1270. doi:DOI 10.1098/rstb.1998.0281

Hintzman, D. L. (1986). Schema Abstraction in a Multiple-Trace Memory Model. *Psychological Review, 93*(4), 411-428. doi:Doi 10.1037/0033-295x.93.4.411

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience, 3*(1), 80-84. doi:Doi 10.1038/71152

Horovitz, S. G., Rossion, B., Skudlarski, P., & Gore, J. C. (2004). Parametric design and correlational analyses help integrating fMRI and electrophysiological data during face processing. *Neuroimage, 22*(4), 1587-1595. doi:10.1016/j.neuroimage.2004.04.018

Hubel, D. H., & Wiesel, T. N. (1998). Early exploration of the visual cortex. *Neuron, 20*(3), 401-412. doi:Doi 10.1016/S0896-6273(00)80984-8

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron, 76*(6), 1210-1224. doi:10.1016/j.neuron.2012.10.014

Ibrahimagić-Šeper, L., Čelebić, A., Petričević, N., & Selimović, E. (2006). Anthropometric differences between males and females in face dimensions and dimensions of central maxillary incisors. *Medicinski glasnik, 3* (2), 58-62.

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping, 38*(3), 1541-1573. doi:10.1002/hbm.23471

Ince, R. A., Jaworska, K., Gross, J., Panzeri, S., van Rijsbergen, N. J., Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*. doi:10.1093/cercor/bhw196

Ince, R. A., van Rijsbergen, N. J., Thut, G., Rousselet, G. A., Gross, J., Panzeri, S., & Schyns, P. G. (2015). Tracing the Flow of Perceptual Features in an Algorithmic Brain Network. *Sci Rep, 5*, 17681. doi:10.1038/srep17681

Izard, C. E. (1994). Innate and Universal Facial Expressions - Evidence from Developmental and Cross-Cultural Research. *Psychological Bulletin, 115*(2), 288-299. doi:Doi 10.1037//0033-2909.115.2.288

Jack, R. E. (2013). Culture and facial expressions of emotion. *Visual Cognition, 21*(9-10), 1248-1286. doi:10.1080/13506285.2013.835367

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Curr Biol, 19*(18), 1543-1548.doi:10.1016/j.cub.2009.07.051

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *J Exp Psychol Gen, 141*(1), 19-25.doi:10.1037/a0023463

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr Biol, 24*(2), 187-192.doi:10.1016/j.cub.2013.11.064

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proc Natl Acad Sci U S A, 109*(19), 7241-7244.doi:10.1073/pnas.1200155109

Jack, R. E., & Schyns, P. G. (2015). The Human Face as a Dynamic Tool for Social Communication. *Curr Biol, 25*(14), R621-634. doi:10.1016/j.cub.2015.05.052

Jack, R. E., & Schyns, P. G. (2017). Toward a Social Psychophysics of Face Communication. *Annu Rev Psychol, 68*, 269-297. doi:10.1146/annurev-psych-010416-044242

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

Johannes, S., Munte, T. F., Heinze, H. J., & Mangun, G. R. (1995). Luminance and spatial attention effects on early visual processing. *Brain Res Cogn Brain Res, 2*(3), 189-205.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol, 58*(6), 1233-1258. doi:10.1152/jn.1987.58.6.1233

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception, 30*(7), 875-887. doi:DOI 10.1068/p3131

Kanade, T., Cohn, J. F., & Tian, Y. (2000). *Comprehensive database for facial expression analysis.* Paper presented at the Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Los Alamitos, CA.

Kanwisher, N., Chun, M. M., McDermott, J., & Ledden, P. J. (1996). Functional imaging of human visual recognition. *Brain Res Cogn Brain Res, 5*(1-2), 55-67.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci, 17*(11), 4302-4311.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352-U357. doi:10.1038/nature06713

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L. Z., & Pascalis, O. (2007). The other-race effect develops during infancy - Evidence of perceptual narrowing. *Psychological Science, 18*(12), 1084-1089. doi:DOI 10.1111/j.1467-9280.2007.02029.x

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. doi:DOI 10.1146/annurev.psych.55.090902.142005

Kim, J. N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience, 2*(2), 176-185.

Kornblith, S., Cheng, X., Ohayon, S., & Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron, 79*(4), 766-781. doi:10.1016/j.neuron.2013.06.015

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science, 293*(5534), 1506-1509. doi:DOI10.1126/science.1061133

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron, 93*(3), 480-490. doi:10.1016/j.neuron.2016.12.041

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn Sci, 17*(1), 26-49. doi:10.1016/j.tics.2012.10.011

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science, Vol 1, 1*, 417-446. doi:10.1146/annurev-vision-082114-035447

Krumhuber, E., & Kappas, A. (2005). Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior, 29*(1), 3-24. doi:10.1007/s10919-004-0887-x

Kruschke, J. K. (2008). Models of Categorization. *Cambridge Handbook of Computational Psychology*, 267-301. doi:Book_Doi 10.1017/Cbo9780511816772

Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2014). A conceptual framework of computations in mid-level vision. *Front Comput Neurosci, 8*, 158. doi:10.3389/fncom.2014.00158

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24*(8), 1377-1388. doi:Pii 93002027510.1080/02699930903485076

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi:10.1038/nature14539

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791. doi:10.1038/44565

Lee, H., & Kuhl, B. A. (2016). Reconstructing Perceived and Retrieved Faces from Activity Patterns in Lateral Parietal Cortex. *J Neurosci, 36*(22), 6069-6082. doi:10.1523/JNEUROSCI.4286-15.2016

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America a-Optics Image Science and Vision, 20*(7), 1434-1448. doi:Doi 10.1364/Josaa.20.001434

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), 572-575. doi:10.1038/nature-04951

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. doi:10.1038/82947

Link, S. W., & Heath, R. A. (1975). Sequential Theory of Psychological Discrimination. *Psychometrika, 40*(1), 77-105. doi:Doi 10.1007/Bf02291481

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience, 5*(9), 910-916. doi:10.1038/nn909

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 24*(3), 732-753. doi:Doi 10.1037/0278-7393.24.3.732

Luck, S. J. (1995). Multiple mechanisms of visual-spatial attention: Recent evidence from human electrophysiology. *Behavioural Brain Research, 71*(1-2), 113-123. doi:Doi 10.1016/0166-4328(95)00041-0

Luck, S. J., & Hillyard, S. A. (1995). The role of attention in feature detection and conjunction discrimination: an electrophysiological analysis. *Int J Neurosci, 80*(1-4), 281-297.

Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998). The Japanese female facial expression (JAFFE) database.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., . . . Tootell, R. B. H. (1995). Object-Related Activity Revealed by Functional Magnetic-Resonance-Imaging in Human Occipital Cortex. *Proceedings of the National Academy of Sciences of the United States of America, 92*(18), 8135-8139. doi:DOI 10.1073/pnas.92.18.8135

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco: W.H. Freeman.

Mcclelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception .1. An Account of Basic Findings. *Psychological Review, 88*(5), 375-407. doi:Doi 10.1037//0033-295x.88.5.375

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 17063-17068. doi:10.1073/pnas.1116333109

Medin, D. L., & Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review, 85*(3), 207-238. doi:Doi 10.1037//0033-295x.85.3.207

Megevand, P., Groppe, D. M., Goldfinger, M. S., Hwang, S. T., Kingsley, P. B., Davidesco, I., & Mehta, A. D. (2014). Seeing Scenes: Topographic Visual Hallucinations Evoked by Direct Electrical Stimulation of the Parahippocampal Place Area. *Journal of Neuroscience, 34*(16), 5399-5405. doi:10.1523/Jneurosci.5202-13.2014

Mumford, D. (1992). On the Computational Architecture of the Neocortex .2. The Role of Corticocortical Loops. *Biological Cybernetics, 66*(3), 241-251. doi:Doi 10.1007/Bf00198477

Murase, H., & Nayar, S. K. (1995). Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision, 14*(1), 5-24. doi:Doi10.1007/Bf0142-1486

Murray, R. F. (2011). Classification images: A review. *J Vis, 11*(5). doi:10.1167/11.5.2

Nasr, S., Liu, N., Devaney, K. J., Yue, X., Rajimehr, R., Ungerleider, L. G., & Tootell, R. B. (2011). Scene-selective cortical regions in human and nonhuman primates. *J Neurosci, 31*(39), 13771-13785. doi:10.1523/JNEUROSCI.2792-11.2011

Nelson, N. L., & Russell, J. A. (2013). Universality Revisited. *Emotion Review, 5*(1), 8-15. doi:10.1177/1754073912457227

Nestor, A., Plaut, D. C., & Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. *Proc Natl Acad Sci U S A, 113*(2), 416-421. doi:10.1073/pnas.1514551112

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, 427-436.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15*(1), 1-25. doi:DOI 10.1002/hbm.1058

Niemeier, M., Goltz, H. C., Kuchinad, A., Tweed, D. B., & Vilis, T. (2005). A contralateral preference in the lateral occipital area: sensory and attentional mechanisms. *Cerebral Cortex, 15*(3), 325-331. doi:10.1093/cercor/bhh134

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen, 115*(1), 39-61.

O'Toole, A. J. (2011). Cognitive and Computational Approaches to Face Recognition In G. Rhodes, A. Calder, M. Johnson, & J. V. Haxby (Eds.), *The Oxford Handbook of Face Perception* (pp. 15 -30).

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends Cogn Sci, 22*(9), 794 - 809. doi:10.1016/j.tics.2018.06.006

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers & generative models. *Cognitive Science, 28*(2), 227-239. doi:10.1016/j.cogsci.2004.09.004

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*. doi:Artn 15686910.1155/2011/156869

Op de Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Reviews Neuroscience, 9*(2), 123-135. doi:10.1038/nrn2314

Palmeri, T. J., Wong, A. C. N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences, 8*(8), 378-386. doi:DOI 10.1016/j.tics.2004.06.001

Panzeri, S., Magri, C., & Logothetis, N. K. (2008). On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magnetic Resonance Imaging, 26*(7), 1015-1025. doi:10.1016/j.mri.2008.02.019

Pasupathy, A., & Connor, C. E. (1999). Responses to contour features in macaque area V4. *J Neurophysiol, 82*(5), 2490-2502. doi:10.1152/jn.1999.82.5.2490

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol, 86*(5), 2505-2519. doi:10.1152/jn.2001.86.5.2505

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience, 5*(12), 1332-1338. doi:10.1038/nn972

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *J Vis, 15*(7), 5. doi:10.1167/15.7.5

Philiastides, M. G., & Heekeren, H. R. (2009). Spatiotemporal characteristics of perceptual decision making in the human brain. *Handbook of Reward and Decision Making*, 185-212. doi:Doi 10.1016/B978-0-12-374620-7.00008-X

Phillips, P. J., Hill, M. Q., Swindle, J. A., & O'Toole, A. J. (2015). Human and Algorithm Performance on the PaSC Face Recognition Challenge. *2015 Ieee 7th International Conference on Biometrics Theory, Applications and Systems (Btas 2015)*.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74-85. doi:10.1016-/j.imavis.2013.12.002

Phong, B. T. (1975). Illumination for Computer Generated Pictures. *Communications of the Acm, 18*(6), 311-317. doi:Doi 10.1145/360825.360839

Pitcher, D., Walsh, V., & Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research, 209*(4), 481-493. doi:10.1007/s00221-011-2579-1

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Curr Biol, 17*(18), 1568-1573. doi:10.1016/j.cub.2007.07.063

Ploran, E. J., Nelson, S. M., Velanova, K., Donaldson, D. I., Petersen, S. E., & Wheeler, M. E. (2007). Evidence accumulation and the moment of recognition: Dissociating perceptual

recognition processes using fMRI. *Journal of Neuroscience, 27*(44), 11912-11924. doi:10.1523/Jneurosci.3522-07.2007

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the United States of America, 115*(15), 3972-3977. doi:10.1073/pnas.1716090115

Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research, 39*(11), 1963-1989. doi:Doi 10.1016/S0042-6989(98)00279-X

Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review, 85*(2), 59-108. doi:Doi10.1-037//0033-295x.85.2.59

Reed, S. K. (1972). Pattern Recognition and Categorization. *Cognitive Psychology, 3*(3), 382-407. doi:Doi 10.1016/0010-0285(72)90014-X

Rhodes, G., Brake, S., & Atkinson, A. P. (1993). What's lost in inverted faces? *Cognition, 47*(1), 25-57.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research, 46*(18), 2977-2987. doi:10.1016/j.visres.2006.03.002

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.

Rodriguez-Sanchez, A. J., & Tsotsos, J. K. (2012). The Roles of Endstopped and Curvature Tuned Computations in a Hierarchical Representation of 2D Shape. *PLoS One, 7*(8). doi:ARTN e4205810.1371/journal.pone.0042058

Rosch, E., & Mervis, C. B. (1975). Family Resemblances - Studies in Internal Structure of Categories. *Cognitive Psychology, 7*(4), 573-605. doi:Doi 10.1016/0010-0285(75)90024-9

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). doi:Artn 310.1167/8.4.3

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport, 11*(1), 69-74.

Rozin, P., & Fallon, A. E. (1987). A Perspective on Disgust. *Psychological Review, 94*(1), 23-41. doi:Doi 10.1037//0033-295x.94.1.23

Rozin, P., Lowery, L., & Ebert, R. (1994). Varieties of Disgust Faces and the Structure of Disgust. *Journal of Personality and Social Psychology, 66*(5), 870-881. doi:Doi 10.1037//0022-3514.66.5.870

Rugg, M. D., Milner, A. D., Lines, C. R., & Phalp, R. (1987). Modulation of Visual Event-Related Potentials by Spatial and Nonspatial Visual Selective Attention. *Neuropsychologia, 25*(1a), 85-96. doi:Doi 10.1016/0028-3932(87)90045-5

Russell, J. A. (1993). Forced-Choice Response Format in the Study of Facial Expression. *Motivation and Emotion, 17*(1), 41-51. doi:Doi 10.1007/Bf00995206

Russell, J. A. (1994). Is There Universal Recognition of Emotion from Facial Expression - a Review of the Cross-Cultural Studies. *Psychological Bulletin, 115*(1), 102-141. doi:Doi10.1037/00-33-2909.115.1.102

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience, 23*(37), 11539-11553.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*(8), 819-825. doi:Doi 10.1038/90526

Schwiedrzik, C. M., & Freiwald, W. A. (2017). High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron, 96*(1), 89-97 e84. doi:10.1016/j.neuron.2017.09.007

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147-179. doi:Doi 10.1016/S0010-0277(98)00016-X

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychol Sci, 13*(5), 402-409.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behav Brain Sci, 21*(1), 1-17; discussion 17-54.

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of brain responses. *Neuroreport, 14*(13), 1665-1669. doi:10.1097/01.wnr.0000088408.04452.e9

Schyns, P. G., Petro, L. S., & Smith, M. L. (2007). Dynamics of visual information integration in the brain for categorizing facial expressions. *Curr Biol, 17*(18), 1580-1585. doi:10.1016/j.cub.2007.08.048

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology-Learning Memory and Cognition, 23*(3), 681-696. doi:Doi10.1037/0278-7393.23.3.681

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A, 104*(15), 6424-6429. doi:10.1073/pnas.0700622104

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379-423. doi:DOI 10.1002/j.1538-7305.1948.tb01338.x

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*(6869), 318-320. doi:10.1038/415318a

Slotnick, S. D., & Schacter, D. L. (2004). A sensory signature that distinguishes true from false memories. *Nature Neuroscience, 7*(6), 664-672. doi:10.1038/nn1252

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences of the United States of America, 107*(46), 20099-20103. doi:10.1073/pnas.1000233107

Smith, M. L., Gosselin, F., & Schyns, P. G. (2012). Measuring Internal Representations from Behavioral and Brain Data. *Current Biology, 22*(3), 191-196. doi:10.1016/j.cub.2011.11.061

Soto, F. A., & Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition, 139*, 105-129. doi:10.1016/j.cognition.2015.02.006

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience, 15*(12).

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nature Neuroscience, 11*(7), 843-850. doi:10.1038/nn.2138

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science, 16*(2), 145-151. doi:DOI 10.1111/j.0956-7976.2005.00795.x

Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 46*(2), 225-245. doi:Doi 10.1080/14640749308401045

Tanaka, J. W., & Taylor, M. (1991). Object Categories and Expertise - Is the Basic Level in the Eye of the Beholder. *Cognitive Psychology, 23*(3), 457-482. doi:Doi 10.1016/00100285(91)90016-H

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research, 51*(11), 1273-1278. doi:10.1016/j.visres.2011.04.002

Tenebaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629-+.

Timme, N., Alford, W., Flecker, B., & Beggs, J. M. (2014). Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of Computational Neuroscience, 36*(2), 119-140. doi:10.1007/s10827-013-0458-4

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cogn Psychol, 12*(1), 97-136.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*, 411-437. doi:10.1146/annurev.neuro.30.051606.094238

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *J Cogn Neurosci, 3*(1), 71-86. doi:10.1162/jocn.1991.3.1.71

Tversky, A. (1977). Features of Similarity. *Psychological Review, 84*(4), 327-352. doi:Doi 10.1037/0033-295x.84.4.327

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11*(2), 58-64. doi:10.1016/j.tics.2006.11.009

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550-592. doi:10.1037//0033295x.108.3.550

Valdes-Sosa, M., Bobes, M. A., Rodriguez, V., & Pinilla, T. (1998). Switching attention without shifting the spotlight: Object-based attentional modulation of brain potentials. *Journal of Cognitive Neuroscience, 10*(1), 137-151. doi:Doi 10.1162/089892998563743

Valentine, T., & Bruce, V. (1986). The Effect of Race, Inversion and Encoding Activity Upon Face Recognition. *Acta Psychologica, 61*(3), 259-273. doi:Doi 10.1016/0001-6918(86)90085-5

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion, 11*(4), 907-920. doi:10.1037/a0023853

Van Essen, D. C., Anderson, C. H., & Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science, 255*(5043), 419-423.

van Rijsbergen, N., Jaworska, K., Rousselet, G. A., & Schyns, P. G. (2014). With age comes representational wisdom in social signals. *Curr Biol, 24*(23), 2792-2796. doi:10.1016/j.cub.2014.09.075

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci, 13*(4), 454-461.

VanVeen, B. D., vanDrongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *Ieee Transactions on Biomedical Engineering, 44*(9), 867-880. doi:Doi 10.1109/10.623056

Weiner, K. S., & Grill-Spector, K. (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *Neuroimage, 52*(4), 1559-1573. doi:10.1016/j.neuroimage.2010.04.262

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proc Biol Sci, 282*(1814). doi:10.1098/rspb.2015.1292

Wijers, A. A., Lange, J. J., Mulder, G., & Mulder, L. J. M. (1997). An ERP study of visual spatial attention and letter target detection for isoluminant and nonisoluminant stimuli. *Psychophysiology, 34*(5), 553-565. doi:DOI 10.1111/j.1469-8986.1997.tb01742.x

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology, 92*(3), 1830-1839. doi:10.1152/jn.00155.2004

Wolfe, J. M. (1994). Guided Search 2.0 - a Revised Model of Visual-Search. *Psychonomic Bulletin & Review, 1*(2), 202-238. doi:Doi 10.3758/Bf03200774

Xu, T., Zhan, J., Garrod, O. G. B., Torr, P. H. S., Zhu, S. C., Ince, R. A., & Schyns, P. G. (2018). Deeper Interpretability of Deep Networks. *arXiv*.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience, 11*(11), 1352-1360. doi:10.1038/nn.2202

Yin, L. J., Chen, X. C., Sun, Y., Worm, T., & Reale, M. (2008). A High-Resolution 3D Dynamic Facial Expression Database. *2008 8th Ieee International Conference on Automatic Face & Gesture Recognition (Fg 2008), Vols 1 and 2*, 116-121.

Yin, R. K. (1969). Looking at Upside-down Faces. *Journal of Experimental Psychology, 81*(1), 141-&. doi:DOI 10.1037/h0027474

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

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational Information in Face Perception. *Perception, 16*(6), 747-759. doi:DOI 10.1068/p160747

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics-Uk, 36*(3), 152-162. doi:10.1016/j.cag.2011.12.002

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308. doi:10.1016/j.tics.2006.05.002

Zanto, T. P., Rubens, M. T., Thangavel, A., & Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience, 14*(5), 656-U156. doi:10.1038/nn.2773

Zhan, J., Garrod, O. G., Van Rijsbergen, N., & Schyns, P. G. (2017). Efficient information contents flow down from memory to predict the identity of faces. *bioRxiv*(125591).

Zhan, J., Ince, R. A. A., van Rijsbergen, N., & Schyns, P. G. (2019). Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. *Curr Biol, 29*(2), 319-326 e314. doi:10.1016/j.cub.2018.11.049

Zhu, S. C. M., D. (2007). A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision, 2*(4), 259-362.