UNIVERSITY OF CALIFORNIA

Santa Barbara

Face Perception: The Interaction of Eye Movements with Internal Face Representations

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychological and Brain Sciences

by

Yuliy Tsank

Committee in charge:

Professor Miguel Eckstein, Chair

Dr. Craig Abbey

Professor Barry Giesbrecht

Professor Mary Hegarty

September 2019

The dissertation of Yuliy Tsank is approved.

_____

Mary Hegarty

_____

Barry Giesbrecht

_____

Craig Abbey

_____

Miguel Eckstein, Committee Chair

May 2019

ACKNOWLEDGEMENTS

In Academia:

Advisor: I would like to thank Miguel Eckstein for his years of mentorship and collaboration, which allowed me to develop my skills and appreciation in using rigorous approaches to visual perception research.

Mentors and Friends: I would like to thank Mordechai Juni and Matt Peterson for their help and advice both in science and in life. In my mind, they are the elder statesmen of the Vision and Image Understanding (VIU) Lab, who hold a commensurate amount of wisdom.

Labmates and Friends: I would like to thank Arturo Deza, Puneeth Chakravarthula, Aditya Jonnalagadda, Katie Koehler, Lauren Welbourne, Eamon Caddigan, Steve Mack, Nicole Han, Devi Klein, and Sudhanshu Srivastava for stimulating discussions (both scientific and social) and contributions to projects.

Outside of Academia:

Family: I would like to thank my parents, Stella Tsank and Leo Tsank, and my brother, Oleg Tsank, for everything, including the unselfishness that allowed me to pursue my professional goals.

Friends: I would like to thank Allic Sivaramakrishnan, Hernan Rosas, Michael Liu, Marvin Thielk, William Scott, Steven Wrieden, Chelsea Lonergan, and Stacy Little for providing a social life outside of academia (sort of, since some of them are just in different fields in academia).

VITA
May 2019

EDUCATION

Bachelor of Arts in Molecular and Cell Biology (Neurobiology), University of California, Berkeley, 2008-2012

Doctor of Philosophy in Psychological and Brain Sciences, University of California, Santa Barbara, 2013-2019 (expected)

Teaching Assistant, Psychological and Brain Sciences, University of California, Santa Barbara, 2013-2019 (expected)

PUBLICATIONS

Tsank, Y., & Eckstein, M. P. (2017). Domain Specificity of Oculomotor Learning after Changes in Sensory Processing. *Journal of Neuroscience*, *37*(47), 11469–11484. https://doi.org/10.1523/JNEUROSCI.1208-17.2017

FIELDS OF STUDY

Major Field: Human Visual Perception

Studies in Face Perception with Professor Miguel Eckstein

Studies in Eye Movements with Professor Miguel Eckstein

Studies in Computational Modeling Methods with Professor Miguel Eckstein

ABSTRACT

Face Perception: The Interaction of Eye Movements with Internal Face Representations

Face perception is a ubiquitous perceptual task that most people easily perform many times a day, beginning in early childhood. The process of extracting meaningful information from a face for tasks such as face identification, gender discrimination, or emotion discrimination, involves making eye movements to different parts of a face. It is known that most of the information for such tasks can be extracted after just a single initial eye movement. However, the efficiency with which that information can be used may be modulated by the internal representation of faces that are stored in our brains for specific tasks. This dissertation explores several aspects of the interaction of the initial eye movement to a face with internal face representations. One aspect is the evaluation of configural processing of faces in the context of a foveated visual system. Another aspect is the effect of natural statistics of facial expressions on the availability of task-relevant information that may be extracted with an initial eye movement to a face. A third aspect is how individual differences in the initial eye movement may shape the development of internal fixation-specific face representations. All of these aspects are investigated using a combination of human psychophysics studies and computational modeling of face perception and eye movements. The data obtained supports a view of the initial eye movement to a face as a highly practiced and consistent behavior that depends on the statistics of the faces that humans are exposed to during different face discrimination tasks. In turn, this behavior may also shape the internal representations of faces in our brains.

TABLE OF CONTENTS

1 Overview

1.1 Introduction

This dissertation is focused on increasing our understanding of the interaction of the initial eye movement to a face with the internal representations of faces that are stored in higher-level brain areas. In the following section, we present a general background of the literature on eye movements to faces. Then we present a summary of three projects that are described in detail in the following chapters: The first project focuses on the effects of having a foveated visual system on configural processing of faces, using a more ecologically valid stimulus set with dynamic facial expressions. The second project focuses on the effects of natural statistics of facial expressions on the initial eye movement to a face. Finally, the third project focuses on how individual differences in the initial eye movement may shape the development of internal fixation-specific face representations.

1.2 Background: Eye Movements to Faces

Face perception and categorization is an important visual ability that humans use many times a day. Most people find it very easy to quickly extract meaningful information from a face to use for a number of possible tasks. These tasks vary in their complexity, ranging from simple gender discrimination to the correct identification of complex emotions such as surprise or anger. All of these tasks, however, require the correct perception of multiple features of the face, with certain features being more important for specific tasks.

Many of these tasks are performed at a conversational distance where the face takes up a large part of an observer's visual field. Due to the variable density of photoreceptors in our retinas and corresponding cortical magnification of more dense regions, humans are unable to process an entire face at a high resolution at this distance. As a result, they must make eye movements in order to focus the high processing power of the foveal region, which is contained in the center of the retina, to specific parts of the face. In order to extract information correctly, saccades must be made to areas of the face that are relevant for the task at hand.

A considerable amount of research has been done on face perception, including research on which parts of a face are used for specific tasks, which parts of a face are targeted by eye movements, and which brain areas are involved in face-related processing (Tsao & Livingstone, 2008). Very little research, however, deals with the functional importance of eye movements to faces and their neural correlates. In addition, most face research involves the use of static face images with a fixed facial expression. Until recently there was little understanding of why humans fixate specific parts of faces when making eye movements. However, recent efforts have concentrated on the functional importance of the

initial fixation to a face (Peterson and Eckstein, 2012).

As a result of the foveation of their visual systems, humans move their eyes in order to focus the high processing power of their fovea onto parts of the visual environment that contain the necessary information for a specific task. In a face discrimination task, those parts of the environment are different features of a face. Previous studies have shown which areas of the face are used by human observers to complete various face discrimination tasks (i.e. gender discrimination vs identity) (Schyns, Bonnar, & Gosselin, 2002; Smith, Cottrell, Gosselin, & Schyns, 2005). For identification, for example, the eyes are the most informative features, followed by the nose and mouth. Although these studies were important in showing which information in a face may be used for different tasks, they did not answer the question of how it was gathered with eye movements.

Many studies have, however, directly recorded the spatial distributions of eye movements to faces for various tasks using static images of faces. One of those studies showed that eye movement behavior in face discrimination tasks was related to different face stimuli in a meaningful way with an "eye-movement based memory-effect," where several saccade characteristics varied between famous and non-famous face stimuli (Althoff & Cohen, 1999). Another study went further in showing that eye-movements are functional and related to performance in face discrimination tasks by showing that there were deficits in performance of recognizing learned faces when fixations were restricted as opposed to being allowed to make eye movements during a learning stage (Henderson, Williams, & Falk, 2005). Several studies, including the two mentioned above had also found that the spatial distribution of eye movements to faces followed a ubiquitous "T" pattern when averaged across observers, with most fixations being made between the eyes and nose, followed by the

mouth, which implied that internal face features were important in various face discrimination tasks (Althoff & Cohen, 1999; Barton, Radcliffe, Cherkasova, Edelman, & Intriligator, 2006; Henderson et al., 2005; Walker-Smith, Gale, & Findlay, 1977).

Although there are many similarities in the overall spatial distribution of eye movements to faces for various tasks when averaging across them, other studies have observed differences in saccade behavior between different populations, individuals, and different tasks as well as stimuli. One study found that those with ASD (Autism Spectrum Disorder) looked at the eyes less than control individuals, but only when viewing complex emotions (Rutherford & Towns, 2008). Since ASD individuals are known to be less perceptive of complex emotions, it is fitting that they would fixate less on the eye region, which is most informative for many complex emotion discrimination tasks (M. F. Peterson & Eckstein, 2012; Schyns et al., 2002). Other studies found individual differences in the spatial distribution of eye movements between healthy individuals when viewing the same set of stimuli (Kanan, Bseiso, Ray, Hsiao, & Cottrell, 2015; Mehoudar, Arizpe, Baker, & Yovel, 2014a; M. F. Peterson & Eckstein, 2013). These individual-specific eye movements remained consistent over time periods that spanned months, suggesting the programming of scanpaths to faces that may be a function of differences in the visual system or encoding of faces between observers. Taken together, these results suggest that that eye movements are made to some internal representation of a face template, which supports an large body of literature on holistic processing in faces (Farah, Wilson, Drain, & Tanaka, 1998).

Studies that have focused on identification performance have found that it only takes as little as one to two fixations to a face to get a high degree of accuracy (Hsiao & Cottrell, 2008; M. F. Peterson & Eckstein, 2012). The study by (M. F. Peterson & Eckstein, 2012)

was the first to rigorously study the critical first eye movement to a face using an ideal observer analysis. Depending on the level of noise or uncertainty that exists, both in the environment and inside the observer, the first eye movement can be sufficient to extract a large amount of information from a face. In the projects presented in this dissertation, we expand on this research of functional eye movements to understand how different fixation positions interact with internal face representations in the visual cortex.

1.3 Project Summaries

The role of eye movements and configural representations in face processing.

Previous research has shown that performance in human face discrimination tasks can be degraded by manipulating the position of features (i.e. eyes, nose, mouth) within a face stimulus (Civile et al., 2018; Collishaw & Hole, 2000; Tanaka & Farah, 1993). This performance difference is typically attributed to a disruption of face mechanisms in higher-order visual areas of the brain involving feature configurations. In this project, we investigate the possibility of the limitation of foveated processing contributing to this performance difference by partitioning performance differences resulting from scrambling face features, into different causes: 1) The proximity of informative features to an optimal point of fixation; 2) Suboptimal fixation strategies; 3) Configural representations in the brain. We use computational models to isolate different aspects of face perception along the visual stream and compare the behavior of the models with those isolated aspects to human fixation behavior and performance in an emotion discrimination ask. We conclude that the vast majority of the magnitude of performance differences across different face configurations may be attributed to configural face mechanisms.

Eye movements during gender discrimination of faces are adapted to the naturally occurring statistics of emotional expressions.

The human visual system programs eye movements for specific tasks by taking into account both the varying resolution of the retina and the distribution of visual task-relevant statistical regularities. Face perception tasks are heavily practiced and involve a very consistent location of important face features, which direct the first eye movement to a

performance-maximizing optimal point of fixation below the eyes (M. F. Peterson & Eckstein, 2012). However, it is unknown to what extent humans use even more fine-tuned statistical properties, like facial expression frequencies during specific face discrimination tasks to adapt their initial eye movement accordingly. In this project, we run a face perception task with an unusual statistical frequency of facial expressions that contain extra information at new spatial locations for specific tasks. However, we show that humans are unable to take advantage of this new information, even when forced to fixate at a new theoretical optimal point of fixation, and do not adjust their initial eye movement. Our results suggest that observers learn an optimal point of fixation to faces using the natural statistics of occurrence of facial expressions for specific tasks and are inflexible to greatly altered facial expression statistics.

The development of internal fixation-specific face representations.

Previous research has shown that humans have a preferred initial fixation position to faces during common face discrimination tasks. However, there are individual differences in the location of this preferred point across observers (M. F. Peterson & Eckstein, 2013). In addition, there are differences in observers' empirical optimal points of fixation, such that observes maximize their performance in a face identification task when forced to fixate closest to their individual preferred fixation location. In this project, we divide a set of observers into two groups based on their preferred fixation location: "eye-lookers" and "nose-lookers." We then test two hypotheses that attempt to explain what causes the individual differences between these two groups of observers. The first hypothesis involves differences in general low-level vision, between eye-lookers and nose-lookers. The second

hypothesis involves differences in internal fixation-specific face representations between eye-lookers and nose-lookers. We implement these internal representations with computational models that are able to represent important differences in human performance between eye-lookers and nose-lookers, when they are forced to fixate different locations along the vertical midline of the face. Our results suggest that individual differences in fixation position between observers are face-specific rather than a more general difference in low-level vision. Our modeling efforts provide evidence that these face-specific differences involve fixation-specific representations in the brain.

2 The Role of Eye Movements and Configural Representations in Face Processing

2.1 Introduction

Face perception is a ubiquitous task that most humans perform many times a day. Many studies have focused on an aspect of face processing that is thought to involve holistic or configural representations of face stimuli in higher-order visual areas. The umbrella terms "holistic" or "configural" processing refer to information that is encoded about the relationship between face features, such as the angles, relative distances, and relative sizes of the features to each other. This is different from information contained in individual features themselves, like the shape and details within the eyes, mouth, nose, etc. This aspect of face processing is thought to be unique to faces and either does not exist at all, or exists to a much lesser extent, in the processing of other complex objects. There have been several extensively used paradigms to study holistic processing, that involve manipulating either the position, or orientation of features of a face (Farah, Tanaka, & Drain, 1995; Farah, Wilson, Drain, & Tanaka, 1998; Tanaka & Farah, 1993; R. K. Yin, 1969; Young, Hellawell, & Hay, 1987). Such manipulations generally affect human performance in various face discrimination tasks under time constraints, or affect reaction times when the response time is unlimited. There is also disagreement about what exactly holistic face processing means, whether it is a different

concept than configural processing, which task to use when studying it (see (Richler, Palmeri, & Gauthier, 2012) for a review), and what causes differences in processing between faces and objects in the first place (Bukach, Bub, Gauthier, & Tarr, 2006; McKone, Kanwisher, & Duchaine, 2007). Regardless of the disagreement in what mechanisms may explain differences in performance that are observed when face features are manipulated in some way, all current theories involve different aspects of processing in higher-order visual areas (Farah et al., 1998; Richler, Gauthier, Wenger, & Palmeri, 2008). However, there may also be other low-level aspects of the visual system early in the visual processing stream that affect performance in face discrimination tasks with altered face features.

One such low-level visual aspect is an interaction between the spatial configuration of the facial features and the varying quality of spatial processing across the human visual field imposed by the foveated visual system. The brain programs eye movements by taking into account the foveated properties of the visual system (G. E. Legge, Klitz, & Tjan, 1997; Gordon E. Legge, Hooven, Klitz, Stephen Mansfield, & Tjan, 2002) in conjunction with the distribution of task-relevant information in the environment (Hayhoe & Ballard, 2005, 2014) to maximize the acquisition of information during basic perceptual tasks using optimal (Najemnik & Geisler, 2005) or heuristic (Morvan & Maloney, 2012; Najemnik & Geisler, 2009; Paulun, Schütz, Michel, Geisler, & Gegenfurtner, 2015) oculomotor strategies. In the context of face perception, humans direct their initial gaze towards a featureless point just below the eyes (Or, Peterson, & Eckstein, 2015; Peterson & Eckstein, 2012, 2013; Peterson, Lin, Zaun, & Kanwisher, 2016; Tsank & Eckstein, 2017). This allows them to simultaneously extract relatively high quality information from the various critical features of a face, such as the eyes, nose, and mouth (see Figure 1a for an example of simulated fixation

positions at two different locations). These preferred initial points of fixation are consistent across time within the same observer. Critically, these preferred points of fixation also optimize face identification accuracy. When observers are instructed to fixate away from their preferred point of fixation, their accuracy in identifying faces diminishes (Or et al., 2015; Peterson & Eckstein, 2012, 2013). Furthermore, the initial optimal point of fixation to a face can be theoretically predicted with a computational model that takes into account the varying quality of visual processing across the visual field and the distribution of information across facial features (Foveated Ideal Observer, FIO; (Peterson & Eckstein, 2012)).

The previous research on the initial eye movement to a face shows that it has functional importance in face perception tasks. This motivates the examination of eye movements and foveated vision for their possible contribution to the variation of perceptual performance across altered face configurations. Previous studies have considered the role of eye movements in configural face processing. Several studies focused on the role of eye movements in the Face Inversion Effect, where performance in face tasks is known to drop when a face stimulus is inverted compared to being presented in an upright orientation. One of those studies (Van Belle, De Graef, Verfaillie, Rossion, & Lefèvre, 2010) focused on separating the effects of holistic vs. featural processing by using a gaze-contingent visibility window that allowed participants to use either the center or periphery of their visual field during the task. That study showed that the Face Inversion Effect can be modulated by restricting participants' visibility to only the center of their visual field. Two other studies (Hills, Cooper, & Pake, 2013; Xu & Tanaka, 2013) also showed that there are differences in both the initial eye movement and gaze behavior in general, between viewing upright face stimuli vs. inverted face stimuli, and that the location of gaze can modulate the Face

Inversion Effect. Another study (Bombari, Mast, & Lobmaier, 2009) found differences in scanpaths and gaze duration between disruptions in configural vs. featural representations. Finally, one study (Heering, Rossion, Turati, & Simion, 2008) controlled for gaze behavior during a composite face experiment, which is another common way to measure disruptions in configural processing. In contrast to the studies mentioned above, this study found that gaze behavior is similar during configural manipulations in a composite face experiment and does not contribute to differences in performance. All of the studies mentioned above used some version of a face-matching or face-recognition task and most focused on discerning mechanisms of specific aspects of higher-level configural processing. Here we focus on the contribution of the limitation of a foveated visual system on differences in performance during the disruption of configural face processing in multiple face configurations. We do this while carefully controlling for the possible changes in the amount of information available to observers during a face emotion discrimination task after various configural manipulations. We propose a detailed investigation of possible variations in task-accuracy arising from interactions between the spatial location of features, eye movements, and foveated vision. We use the framework of the foveated Bayesian ideal observer to generate theoretical predictions. These predictions focus on the influences on task-accuracy arising from the interaction between foveated vision and the locations of features in different face configurations. We make empirical measurements of eye movements and task accuracy as a function of fixation to assess the role of eye movements, foveation, and high-level configural representations (often referred to as holistic processing) in performance differences between different face configurations.

There are several sources of information loss related to foveated visual processing

that may contribute to performance differences in face discrimination tasks with altered spatial positions of face features relative to the same tasks with an intact upright face:

1. Variations in foveated task information related to spatial distance of informative facial features:

One possible source of performance variations is that the location of the important features for the task has changed such that they are closer together or further apart from each other relative to where they would be in an intact upright face (Figure 1b.1). In this case, the possible level of performance obtained at the optimal point may be changed. It may either be limited by the inability to simultaneously extract high quality information from important features if they are further apart, or improved by a better ability to do so if they are closer together. The foveated Bayesian ideal observer allows us to make quantitative predictions of the impact on task-accuracy of the variations of spatial distance of informative facial features when face configurations are manipulated.

2. Inability of humans to adapt to new optimal eye movement strategies with unusual face configurations:

A second possible source of performance variation is that there is a change in the optimal point of fixation, but observers are not able to adapt their eye movement strategy to fixate it. In this second scenario, a new unusual face configuration may allow an observer to attain a comparable task accuracy to an intact-face configuration, but requires the observer to learn to fixate the atypical face configuration in a new way. Failure to learn a new optimal point of fixation may result in performance degradation (Figure 1b.2). Previous studies have shown that humans have difficulty adapting to a new optimal fixation position with an unusual set of faces where the only discriminating information between them is a single

feature (Peterson & Eckstein, 2014), or when there is a major change in the visual system (Tsank & Eckstein, 2017). These findings motivate the investigation of an inability to learn new optimal fixation points as a possible source of performance degradation when face configurations are altered.

3. Configural processing disruptions:

The two possible sources of performance variations discussed above, relate low-level disruptions of visual processing. However, another explanation (Figure 1b.3), which is the most common in studies investigating configural effects, is that there is a less efficient extraction of information from altered face features due to either a quantitative or qualitative disruption of face representations stored in higher-level face-specific brain areas (Farah et al., 1998; McKone et al., 2007; Rossion, 2008; Sekuler, Gaspar, Gold, & Bennett, 2004); see (Tsao & Livingstone, 2008) for a review of proposed different stages of face processing. For example, in the qualitative view of the disruption of face processing in the Face Inversion Effect, there is only a small disruption in the processing of featural information (i.e. information contained within a feature, such as the details of an eye), while there is a large disruption in configural information (i.e. information about distances and locations of features relative to each other) (Rossion, 2008). In the quantitative view, however, there is a similar disruption of processing of both featural and configural information (Sekuler et al., 2004). In addition to behavioral studies, there is recent evidence from neurophysiology studies that the location of face features in specific parts of the visual field affects how they are processed. There is evidence for increased neural activity in the posterior lateral face patch in monkeys (Issa & DiCarlo, 2012) and increased separation of activity patterns to major face features in the right inferior occipital gyrus in humans (de Haas et al., 2016),

when the facial features appear at the typical retinal positions relative to a preferred point of fixation, compared to atypical locations. Here, we use comparisons of human and computational models across face configurations as one test for configural representations of faces. The evaluated FIO model has representations of the faces that match any face configuration in the stimuli. As a result, the FIO does not capture any pre-determined preference for any spatial configuration of features. If humans observers have some constraints on the spatial relationships across representations of features, we might expect large divergences in performance between humans and predictions by the FIO. However, in addition to the FIO, we use another computational model that is meant to simulate some of the configural representations that might arise from exposure to the statistical distribution of face configurations through early development. We use a Convolutional Neural Network (CNN) trained from the beginning (i.e. starting with random parameters, rather than a pretrained model) on a face stimuli set with a majority of upright intact faces to simulate this exposure. We then test the model with the different face configurations that we also use for human participants.

To test the hypotheses described above, we measure human eye movements and performance at different fixation vertical positions along presented faces during an emotion discrimination task. We use short movies of faces rather than static images in order to add ecological validity to the stimuli. The task consists of movies of the onset of three possible facial expressions (happiness, sadness, and fear) and five possible face feature configurations, four of which have features that have been rotated or shifted relative to an intact upright face. We ask whether human observers show a performance difference in the task between different face configurations. We use the FIO and a CNN to test the various

hypotheses related to variations in visual information accessible by a foveated visual system, inability of humans to learn how to optimally look at unusual face configurations, and configural representations of faces.

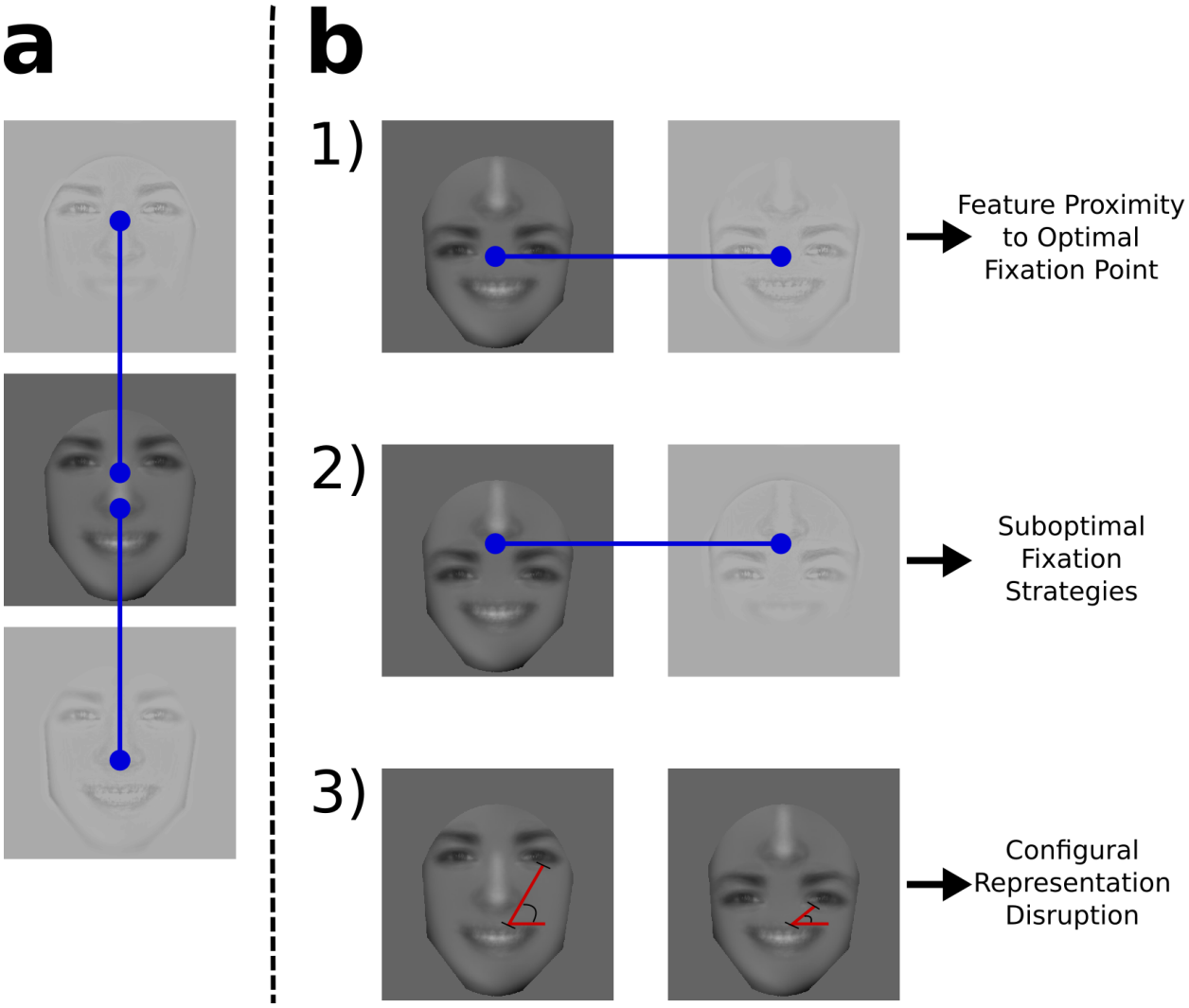

Figure 1: ***a)*** *A standard upright face is shown in the middle with two foveation filters, corresponding to simulated fixation positions at the eyes (top image) and nose (bottom image).* ***b)*** *A sketch of three possible causes of performance disruptions in a face emotion discrimination task are shown. b.1) Here the same face is shown as in part (a), except the position of the eyes and nose has been switched. Now the eyes are closer to the mouth and the nose is further away from the mouth, relative to the original intact face. A simulated fixation position, which is optimal for this task, and is in between all of the internal face features (eyes, nose and mouth), is shown. Performance may change because certain features are closer or further away from the fixation position and a processed with higher or lower resolution, respectively. Depending on which features are important for a particular task, if*

they are processed with lower resolution, this may lead to a decrease in performance. b.2) Here, the same scrambled face stimulus is shown, except with a different simulated fixation position, which is suboptimal. There may be decrease in performance when processing this face stimulus at this fixation position relative to the fixation position shown above. This may happen because the mouth, and partially the eyes, are now processed with lower resolution than before. b.3) Any differences in performance between scrambled faces that cannot be explained by the effects of foveation, will be attributed to a disruption of configural processing in higher-order brain areas. Here, a change in the distance and angle between the eyes and mouth are shown as an example of a difference in configural information between an upright, intact face configuration compared to a face configuration with a switched position of the eyes and nose.

2.2 Materials and Methods

Human Psychophysics Studies.

Participants.

Experiment 1. The first experiment was a free-viewing condition with multiple face configurations and a set presentation time and was completed with a group of eleven undergraduate students of either sex, who participated in the study for course credit. Informed consent was obtained from all subjects and guidelines provided by the institutional review board of the University of California, Santa Barbara were followed.

Experiment 2. The second experiment was a free eye movement and forced-fixation condition with 8 participants. Data from one of the participants was not used because they were found to have a first-fixation strategy (mouth-looker) that is significantly different from the others. Informed consent was obtained from all subjects and guidelines provided by the institutional review board of the University of California, Santa Barbara were followed.

Apparatus and Materials.

MATLAB Psychtoolbox and Eyelinktoolbox software were used to run the eyetracker from a display computer as well as present visual stimuli on the display screen. The display used was a Barco MDRC 1119 monitor set to a 1280x1024 pixel resolution and was located 76.5cm away from the observer's eyes. The display was linearly calibrated with a minimum luminance of .05 cd/m² and a maximum luminance of 126 cd/m².

Eye-tracking.

The left eye of each participant was tracked using an SR Research Eyelink 1000 Tower Mount eye tracker sampling at 250 Hz. A nine-point calibration and validation were run before each 125-trial session, with a mean error of no more than 0.5º of visual angle.

Saccades were classified as events in which eye velocity was greater than 35º and eye acceleration exceeded 9,500º per square second. The recommended thresholds by SR for cognitive research are an eye velocity of 30º and an eye acceleration of 8,000º per square second. The minor increase of the velocity and acceleration thresholds in our parameter settings allowed us to better control the number of "broken fixations" during the initial fixation stage at the beginning of every trial prior to the presentation of the stimulus.

Stimuli.

In these experiments, observers completed an emotion identification task with three emotions: happiness, sadness, and fear. As shown in Figure 2a, the task was done for intact, upright faces, as well as four other configurations of faces where major face features (eyes, nose, mouth) were moved or inverted. As shown in Figure 2b, sixty movies of facial expressions were used for each face configuration, exhibiting the three emotions for each of the same twenty different identities (each identity had a happy, sad, and fearful stimulus). The identities consisted of fifteen female faces and five male Caucasian faces. Each movie contained 1400ms (35 frames with 40ms per frame) of a specific identity starting from a neutral expression and unfolding into one of the three emotions mentioned above. All of the stimuli were originally obtained from (L. Yin, Chen, Sun, Worm, & Reale, 2008), who recorded them from undergraduates at Binghamton University using a 3D face scanner.

Preprocessing. The original set of videos contained 100 frames (4000ms) of six expressions (happiness, disgust, fear, anger, surprise and sadness) from 101 subjects at a resolution of 1040x1329 with front-view poses. The videos were first manually filtered such that those that didn't start with a neutral expression or had too much movement or eye blinking were discarded. We used a subset of twenty subjects and three expressions

(happiness, sadness, and fear) for each subject, after having twenty-five undergraduate participants rate all of the remaining movies for authenticity and intensity of the expression being shown. We chose the subset based on the highest combination of authenticity across the three expressions with a constraint that the same twenty identities be present for all three expressions. The stimuli were first spatially aligned such that the face in the first frame of each movie was positioned with center of the eyes at 2/5 of the image height below the top of the image and with the chin 1/50 of the image height above the bottom of the image. This was done by extracting facial landmarks around the eyes and mouth using the Python dlib library and then rotating, resizing, and cropping the images using the first frame as a reference, and doing the same operations to the rest of the frames that were done to the first frame. The stimuli were then temporally aligned such that the beginning of the facial expression in each set of frames started at approximately the same time. The temporal alignment was done using a normalized cross correlation between the first frame, which contained a neutral facial expression, to each frame in the movie until a threshold was reached at which there was a large enough change in the current frame relative to the first one to signal the start of the facial expression. The frame before the one that crossed the threshold was then chosen as the first frame and contained a neutral facial expression.

Creating Different Face Configurations. In order to create the four face configurations that were used along with the original face configuration, we used a Poisson blending procedure. This was done in order to create blended faces that were as natural-looking as possible. First, individual masks for the eyes, nose, and mouth in each upright, intact face movie were created such that all facial features were inside the boundary of a mask polygon throughout all frames. Then, for each face configuration, a blank facial frame

was created using photoshop, with designated new positions of features. Poisson blending (Pérez, Gangnet, & Blake, 2003) was used to blend facial features from the original face (source) onto the blank facial frame (target for each configuration) by using the corresponding masks for each individual frame of the video.

Stimulus Presentation on Screen. Each frame of each movie was luminance-mean normalized to 62 cd/m² and shown to participants at a Root Mean Square (RMS) contrast of .1517, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 6.83 cd/m² (corresponding to a noise RMS contrast of .11). Participants viewed the face stimuli 76.5cm away from the display resulting in a square stimulus (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. The large size of the faces, more typical of conversational distance, was chosen: (1) to allow measurements of larger variations of perceptual performance with fixation position (for small faces perceptual performance is less sensitive to fixation position within the face); (2) to allow more precise measurements of fixation positions relative to facial features. In addition, the large faces (10 deg. width, 15 deg. height) have been shown to be the face size that optimizes face identification (Yang, Shafai, & Oruc, 2014).

Procedure.

Experiment 1. Observers performed a free-eye movement emotion identification task with 20 total blocks consisting of five different face configurations. Each block consisted of 125 trials. The configurations varied with different positions of features and were presented with two blocks each in the following order: an original intact face, an inverted face, a face with inverted features, a face with the eyes and mouth switched, and a face with the eyes and

nose switched (Figure 1). After a total 10 blocks, the entire presentation order of the five different configurations was repeated again for a total of 20 blocks. The contrast (described in the *Stimulus Presentation on Screen* section above) and stimulus presentation time of 1400ms remained the same for all the blocks.

Experiment 2. Observers performed both a free-eye movement emotion identification task followed by a forced-fixation emotion identification task with 37-39 total blocks consisting of three of the five face configurations from Experiment 1. The configurations included an original intact face, an inverted face, and a face with the eyes and mouth switched. They were presented in separate blocks in the order described and then this block order was repeated 12-13 times. The first three blocks were a free eye-movement condition with a presentation time of 1000ms, while the rest of the blocks were a forced-fixation condition with a presentation time of 200ms. Each block consisted of 125 trials. The contrast (described in the *Stimulus Presentation on Screen* section above) remained the same for all the blocks.

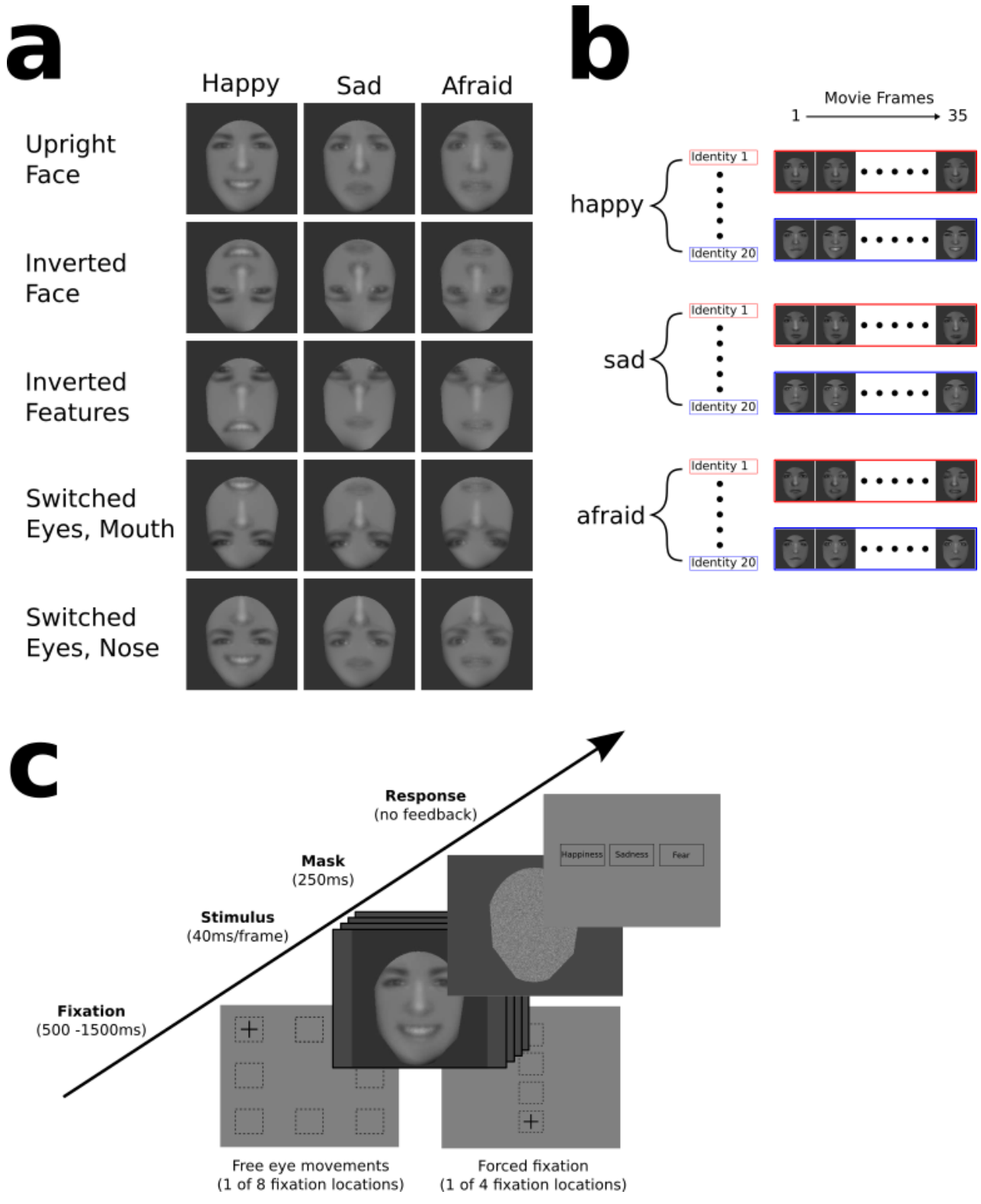

Figure 2: ***a)*** *Averages across face images foreach of the three different emotions and each of the five different configurations. An average image was computed by taking the mean of the grayscale luminance values across 20 individual face images in each configuration and*

emotion category. This was done for illustration purposes. The experiment presented individual faces of each individual. ***b)*** *The stimuli contents of a single configuration are shown in more detail (here the upright face configuration is used as an example). Each configuration contains three emotions, with the same 20 identities represented for each emotion. Each identity has an associated 1400ms (35 frame) long movie. Only parts of each movie are shown in different experiments and conditions.* ***c)*** *A trial time line is shown. In the free-viewing condition of Experiment 1, observers made saccades to the centrally presented face from a fixation cross in one of eight randomly chosen locations. A stimulus presentation time of 1400ms consisting of 35 frames was used in the free-viewing condition. Separate blocks were used for each of the five different configurations, with the three emotions equally intermixed in each block. In the forced-fixation condition of Experiment 2, observers fixated one of four locations on the vertical midline of the face stimuli (approximately corresponding to the forehead, eyes, nose, and mouth). A stimulus presentation time of 200ms was used for the forced-fixation condition. Here an average happy stimulus from the upright face configuration is shown without noise and high contrast for illustration purposes (in the actual experiment contrast was substantially lower and white noise was added). At the end of each trial, participants had unlimited time to select with the mouse one of three emotions that were displayed on the screen.*

Experimental Conditions.

Experiment 1.

Starting Fixation. As shown in the lower left box of Figure 2c, during free-viewing blocks, participants started a trial by pressing the space bar while fixating a cross (.44 º x .44 º) in one of eight randomly chosen locations located on average 13.94º from the center of the stimulus. The fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants moved their eyes more than 1º from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus.

Trial Timing. As shown in the middle of Figure 2c, the stimulus was presented after an initial fixation cross. The stimulus was shown for 1400ms in the free-viewing condition of Experiment 1. At the end of each trial, participants had unlimited time to select with the

mouse one of three emotions that were displayed on the screen. No feedback was given.

Experiment 2.

Starting Fixation. As shown in the lower right box of Figure 2c, during forced-fixation blocks, the cross was located in one of 4 locations, which corresponded to the forehead, eyes, nose, and mouth every 5.07 º downward respectively. The fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants moved their eyes more than 1º from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus.

Trial Timing. As shown in the middle of Figure 2c, the stimulus was presented after an initial fixation cross. The stimulus presentation time was 200ms in the forced-fixation condition of Experiment 2. The short presentation time for the forced-fixation condition was used in order to account for the fact that participants did not need time to make an eye-movement from the periphery of the screen, as they did in the free-viewing condition, as well as to make the task more difficult in order to avoid ceiling effects. In addition, the stimulus in the forced fixation condition was started from the 5th frame (200ms) in order to account for the amount of time it would take to make an eye movement inside the stimulus. At the end of each trial, participants had unlimited time to select with the mouse one of three emotions that were displayed on the screen. No feedback was given.

Ideal Observer Models.

In the following sections, we briefly describe several Ideal Observers that we use to model the face perception task presented in this paper and explain the rest of the algorithmic details in the Appendix. Ideal observer models are not meant to be a biologically realistic representation of face perception. Instead, they measure the overall information available in stimuli that are used in a particular task, when the statistics of the task are fully known. These kinds of models have been used extensively in visual perception research (M. P. Eckstein, Schoonveld, Zhang, Mack, & Akbas, 2015a; Gold et al., 2013; Gold, Mundy, & Tjan, 2012; Najemnik & Geisler, 2005; Sekuler et al., 2004) and are very useful in explaining the mechanisms involved in the way that humans perform specific tasks; see (Geisler, 2011) for a thorough review of what Ideal Observers are and how they are used in different areas of vision research. A standard Ideal Observer acts as a benchmark for the maximum possible performance that can be achieved in a particular task, when optimally using all available information for a task under a specific level of uncertainty. In visual tasks, the uncertainty generally comes from pixel noise added to the stimuli shown, which we call “external noise”, because it is external to an observer. Generally, human performance in most tasks is very far from this benchmark because of various sources of suboptimalities in the human visual system, including noise (uncertainty) in neural firing, suboptimal integration of information across spatial locations, suboptimal encoding of information at various levels in the visual stream, suboptimal biases in the decision stage for particular stimuli, and many more. To explain human performance, specific aspects or a single aspect of a source of suboptimality in the human visual system are added to an ideal observer. If a simulation of an added aspect alone is enough to explain important differences in human behavior and performance, then it

can provide important insight into the mechanisms that are responsible for those differences. In the context of eye movements to faces, a Foveated Ideal Observer (FIO), which we describe below, has previously been used to simulate the foveation of the visual system. Foveation is a suboptimality, relative to a regular Ideal Observer that processes the entire visual field with equally high resolution, and in essence introduces eye movements to an ideal observer. Simulating eye movements in this way, can help determine if the limitation of having a foveated visual system is enough to explain certain important eye movement behaviors. All other suboptimalities in the visual system are added to the FIO model in the form of "internal noise", which brings down performance of the FIO without modeling in detail all of the sources of suboptimality that are meant to be represented by this. The noise is "internal" because it is meant to represent suboptimalities further down the visual pathway and is added to the model after the addition of "external noise as well as after a foveation filter, which processes the noisy incoming stimuli.

Bayesian Ideal Observer.

Here, we run several different variants of an ideal observer model, starting with a standard ideal observer, which utilizes image information to achieve the highest possible performance and does not simulate the foveation of the visual system like the FIO described below. We run a face emotion identification task with a set of 60 (20 of the same identities for each of 3 emotions) front-view facial expression movies that are normalized for the position of the eyes and chin as well as for contrast (see the Stimuli subsection of Human Psychophysics Studies above for details). Each face movie for the ideal observer simulations consists of 5 frames, which matches a 200ms presentation time that was used in the forced-

fixation condition of Experiment 2 (see Trial Timing section of Experimental Conditions above for details). An ideal observer optimally integrates information over time, so for each movie, the 5 frames are concatenated into a single large frame, which effectively treats the time dimension as a spatial dimension. The frames at corresponding times for each face movie now spatially align with frames from the same time period in other movies. On each trial of the simulation, the face movies $\{\mathbf{f}_1,\ldots,\mathbf{f}_{60}\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen. The same contrast and additive white noise that was used for humans is then added to a chosen template, *i*. The input data, $\mathbf{g}$, to the ideal observer on each simulated trial is then the sum of a random (1 of 60) face template, $\mathbf{s}_i$, and external noise, $\mathbf{n}_{ex}$.

$$\mathbf{g} = \mathbf{s}_i + \mathbf{n}_{ex} \qquad (2.1.1)$$

Using Bayes rule, the ideal observer finds a set of posterior probabilities, one for each hypothesis, $H_{e,f}$,that face f from emotion *e* (happy, sad, or afraid) was shown, given the image data, $\mathbf{g}$. Here we use the index, *f*, to represent a calculated posterior probability for a particular face being shown, in contrast to the index, *i*, which represents the actual ground truth signal that was shown on a particular trial.

The posterior probability, $P(H_{e,f}\,|\,\mathbf{g})$, is calculated using the prior probabilities, $P(H_{e,f})$, and the likelihood, $P(\mathbf{g}\,|\,H_{e,f})$, of the image data, $\mathbf{g}$, given the presence of each face, f from emotion *e*:

$$P(H_{e,f}\,|\,\mathbf{g}) = \frac{P(\mathbf{g}\,|\,H_{e,f})P(H_{e,f})}{P(\mathbf{g})} \propto P(H_{e,f})P(\mathbf{g}\,|\,H_{e,f}) = l_f \qquad (2.1.2)$$

Then to find the posterior probability, $P(H_e \mid \mathbf{g})$, of the presence of a specific emotion, the sum is found across the posterior probabilities of individual faces belonging to that emotion:

$$P(H_e \mid \mathbf{g}) = \sum_f P(H_{e,f} \mid \mathbf{g}) \tag{2.1.3}$$

The normalizing factor, $P(\mathbf{g})$, in equation (2.1.2) is the same for all posterior probabilities, so it can be ignored without changing the result. The likelihood, $P(\mathbf{g} \mid H_{e,f})$, of the signal having come from a particular face is calculated from a known distribution that comes from a product of distributions of individual pixel noise (see Appendix for details). The maximum posterior probability is then chosen as the answer at the end of a simulated trial:

$$decision = \underset{e}{\operatorname{argmax}}(P(H_e \mid \mathbf{g})) \tag{2.1.4}$$

Region of Interest Bayesian Ideal Observer.

In order to understand which regions of a face are important for this particular task we also run a Region of Interest Ideal Observer (ROI), which is a Bayesian Ideal Observer that is separately run using small sections of the face stimuli image at a time. We run the ROI for each frame of all the movies separately (i.e. we simulate the emotion discrimination task using only first frame of each movie, then separately the second frame, and so on) in order to see how discriminative information may change as a facial expression develops The calculations are the same as for the ideal observer, except that in contrast to equation (2.1.1), the data, $\mathbf{g}_s$, is now the sum of a random (1 of 60) face template, $\mathbf{s}_{i,s}$, and external noise,

$\mathbf{n}_{ex}$, where *s* indexes the section of the face for which performance is separately calculated:

$$\mathbf{g}_s = \mathbf{s}_{i,s} + \mathbf{n}_{ex} \tag{2.2.1}$$

The signal $\mathbf{s}_{i,s}$ on each simulated trial is now taken from a specific 30x30 pixel section from a randomly chosen face template, *i*. Figure 3a and 3b show how small sections of a face are processed at a time and likelihoods are found for each section. Figure 3c shows a performance map that is created by sampling different sections across the face stimulus. Here, we run a simulation with 30,000 trials. Due to computational constraints, we only sample the sections every 10th pixel rather than every adjacent pixel, which results in a 47x47 performance map (it is not 50x50 because of the 30px section size). This map is then resized using bilinear interpolation to a 500x500 pixel performance map to match the size of the face images.

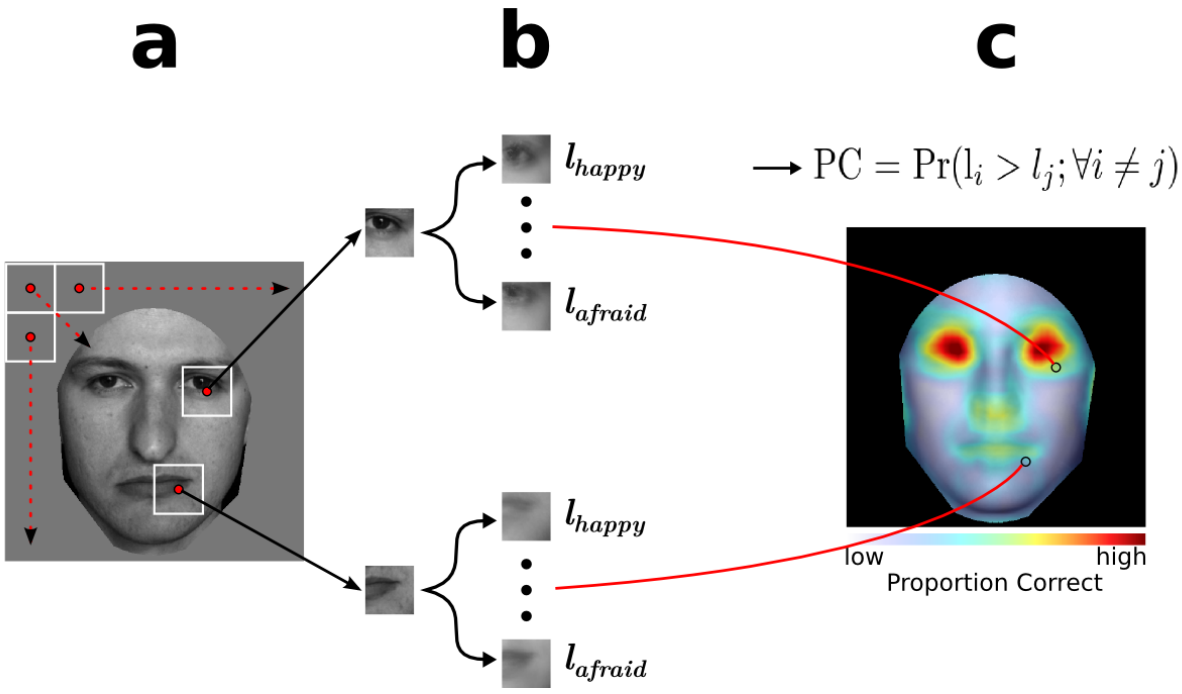

Figure 3: A flow chart for a Region of Interest Ideal Observer. (a) An Ideal Observer is separately run for each small 30x30px section of a face image corresponding to a center point that is sampled every 10px. The ROI is run on a face emotion discrimination task

separately on individual frames of each movie (i.e. we simulate the task using only first frame of each movie, then separately the second frame, and so on). (b) On each simulated trial, likelihoods are found for a chosen face to be a particular emotion, which are found from sums of likelihoods of individual identities representing that emotion. (c) The maximum likelihood principle is used to find performance in the task for each separate face section and output a performance map that shows which parts of a face are the most informative for this task.

Foveated Ideal Observer (FIO) Model.

A spatially variant contrast sensitivity function (SVCSF) was used to model the degradation of the quality of information obtained in the periphery of a foveated visual system (M. F. Peterson & Eckstein, 2012):

$$SVCSF(f,r,\theta) = c_0 f^{a_0} \exp(-b_0 f - d_0(\theta) r^{n_0} f) \quad (2.3.1)$$

where *f* is spatial frequency in cycles per degree of visual angle. The terms a_0, b_0, and c_0, were chosen constants set to 1.2, 0.3, and 0.625 respectively, to set the maximum contrast at 1 and the peak at 4 cycles per degree of visual angle at fixation. The polar coordinates *r* and θ specify the distance in visual angle and direction from fixation. d_0 specifies the eccentricity factor as a function of direction, which represents how quickly information is degraded in the periphery. n_0 specifies the steep eccentricity roll off factor. In the model simulations, different parameters are used for d_0 for the vertical up, *du*, vertical down, *dd*, and horizontal, *dh*, directions. The parameters *du*, *dd* , *dh*, and n_0 are fit to the forced-fixation condition in Experiment 2 in order to match human performance (proportion correct) as a function of fixation position (4 different fixations down the vertical midline of the face) of an emotion discrimination task using upright faces. The values used for parameters *du*, *dd*, *dh*, and n_0 respectively, are 2E-6, 9E-6, 1E-6, and 5. The Akaike Information Criterion (Akaike, 1974),

which takes into account the variance for each data point, is used as a distance measure. The same parameters are used for the emotion discrimination with all other face configurations (see Stimuli section of Human Psychophysics Studies above). Figure 4 shows a flowchart of the algorithmic details of the FIO. The circular plots between Figure 4b.1 and 4b.2 show examples of 2d contrast

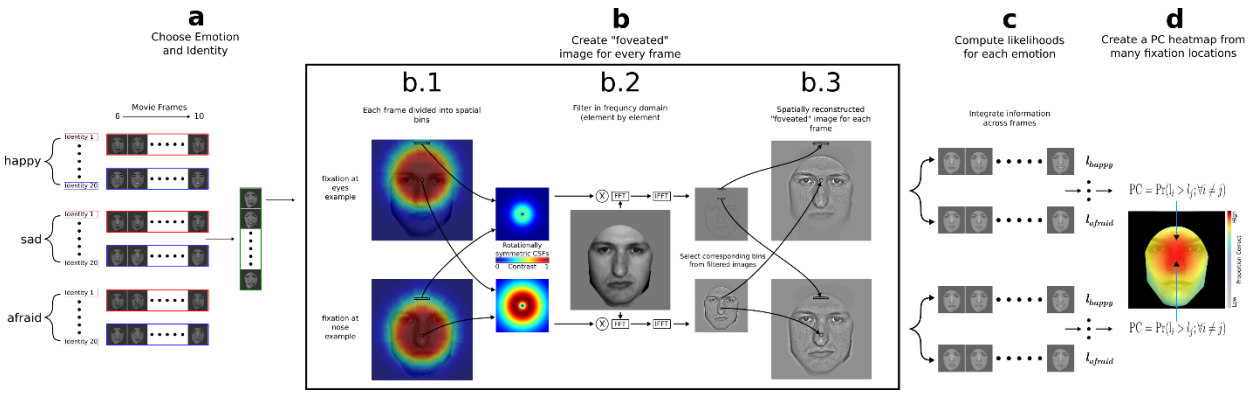

Figure 4: A summary of the process of the computations in the FIO for two fixation positions. The top panels show a fixation point that is below the eyes, which is suboptimal in an emotion discrimination task with upright faces. The bottom panels show a fixation that is above the tip of the nose, which is optimal for this task. ***a)*** *Many trials are simulated where on each trial, a face template is chosen as a signal. Here, the signal selection is shown for an emotion discrimination task for one of 60 face templates, each of which contains 5 frames (because the model was fit to the short-presentation forced-fixation task of Experiment 2).* ***(b.1-b.3)****, The filtering operation for a noiseless template. The filtering operation is done for each frame separately, after which the frames are concatenated together.* ***(b.1)****, A face image is conceptually divided into bins that correspond to specific CSFs as a function of retinal eccentricity. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the CSF plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low-contrast blue in the CSF plots).* ***(b.2)****, The image is transformed into the frequency domain, filtered separately by each possible CSF (here only two are shown), and then transformed back into the spatial domain, resulting in a set of differently filtered images corresponding to each bin.* ***(b.3)****, Corresponding bins are then extracted from the filtered images and input into a composite image that simulates foveation. The procedures in b.1–b.3 are then repeated for each of the frames.* ***c)*** *A set of response variables are then calculated, from which a set of likelihoods is found of each face movie given the noisy image input.* ***d)****, A decision of which face was shown is made by taking the maximum likelihood. Across many trials, a set of proportion correct (PC) values is found, one for each fixation point, and then combined into a heatmap. iFFT, Inverse FFT.*

sensitivity functions at 2 different locations with respect to the fixation position. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low contrast in blue in the plots).

Here, we run a face emotion discrimination task using movies of faces (5 frames long) that start with a neutral expression and develop into one of 3 possible expressions that correspond to happiness, sadness, or fear. We separately run several different conditions where the features of the face stimuli are moved or rotated. We simulate many trials of each condition of each task. On each trial of the simulation, the face templates $\{\mathbf{f}_1,\ldots,\mathbf{f}_n\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen, where *n* is 60 for the face emotion discrimination task (20 identities, with 3 emotions for each identity). Each face template, $\mathbf{f}_i$, consists of 5 changing frames, which is the length of time that was used for stimulus presentation in the forced-fixation condition of Experiment 2. The 5 frames were taken from frames 6-10 out of 35 frames, in order to account for the average amount of time it took human participants to make the first saccade from the periphery of the screen into the face stimulus. The same contrast and additive white noise that was used for psychophysics experiments in humans is then added to a chosen template, *i*, before being linearly filtered with the SVCSF and corrupted with additional internal white noise to become the input data, $\mathbf{g}_k$, to the ideal observer:

$$\mathbf{g}_k = \mathbf{E}_k(\mathbf{s}_i + \mathbf{n}_{ex}) + \mathbf{n}_{in} \qquad (2.3.2)$$

where k, indexes a specific fixation position, $\mathbf{n}_{ex}$ is the external Gaussian white noise, $\mathbf{n}_{in}$ is the internal Gaussian white noise, and $\mathbf{E}_k$ is the linear operator that simulates the fixation dependent foveation of the input. This foveated signal is compared (by taking a dot product) to similarly foveated noiseless templates (original face images) to arrive at a set of responses, $\mathbf{r}_{f,k}$, which come from a multivariate Gaussian distribution with a known mean, $\boldsymbol{\mu}_{f,k}$, and covariance matrix, $\sum_k$ (see Appendix for details on how they are calculated):

$$\mathbf{r}_{f,k} \sim MVN(\boldsymbol{\mu}_{f,k}, \sum_k) \tag{2.3.3}$$

Using Bayes rule, the FIO finds a set of posterior probabilities, one for each hypothesis that face *f* from emotion *e* was shown, $H_{e,f}$, given a set of responses $\mathbf{r}_{f,k}$. The posterior probability, $P(H_{e,f} \mid \mathbf{r}_{f,k})$, is calculated using the prior probabilities, $P(H_{e,f})$, and the likelihood, $P(\mathbf{r}_{f,k} \mid H_{e,f})$, of the set of responses given the presence of each face, *f*, and the observer's fixation at spatial location, *k*:

$$P(H_{e,f} \mid \mathbf{r}_{f,k}) = \frac{P(\mathbf{r}_{f,k} \mid H_{e,f})P(H_{e,f})}{P(\mathbf{r}_{f,k})} \propto P(H_{e,f})P(\mathbf{r}_{f,k} \mid H_{e,f}) \tag{2.3.4}$$

Then to find the posterior probability, $P(H_e \mid \mathbf{r}_{f,k})$, of the presence of a specific emotion, the sum is found across the posterior probabilities of individual faces belonging to that emotion:

$$P(H_e \mid \mathbf{r}_{f,k}) = \sum_f P(H_{e,f} \mid \mathbf{r}_{f,k}) \tag{2.3.5}$$

The maximum posterior probability is then chosen as the answer:

$$decision = \underset{e}{\operatorname{argmax}}(P(H_e \mid \mathbf{r}_{f,k})) \tag{2.3.6}$$

Convolutional Neural Network Model with Fixation-Specific Training and Testing.

Stimuli.

Simulating the statistical distribution of foveated inputs for network training.

The training stimulus set was taken from a combination of the Multimedia Understanding Group (MUG) faces database (Aifanti, Papachristou, & Delopoulos, 2010) and the Cohn-Kanade (CKPlus) faces database (Kanade, Cohn, & Yingli Tian, 2000; Lucey et al., 2010), with permission. Both databases contain movies of facial expressions starting from a neutral expression and unfolding into one of several different emotions, all of which were discarded, except for happiness, sadness, and fear, which were used in the human psychophysics studies, explained above. The databases were combined into a single dataset with movies of 175 male and female participants of various ethnicities that acted out these emotional expressions. Each movie for each emotion was manually trimmed by removing frames from the beginning and end that contain a neutral expression. The frames of each movie were then treated as separate images and were spatially aligned and mean luminance normalized in the same way as the stimuli for the psychophysics studies (see Stimuli section in Human Psychophysics Studies) above, resulting in 1576, 2229, and 1646 individual training images for fearful, happy, and sad expressions, respectively. Multiple copies of the frames were then made by adding a simulation of the foveation of the visual system centered at different fixation positions as shown in Figure 5a below. The fixation positions that were used as the center of foveation (bottom left image of Figure 5a) were taken from empirical fixation positions for upright faces measured in humans in Experiment 1. However, creating a set of face stimuli that were foveated at every possible position in the image would have

required making thousands of copies of the base dataset. This would have made such an implementation intractable with our limited computational resources, both in terms of the amount of time it would take to produce those multiple foveated copies and in the amount of time it would take to train a CNN on all of them. Instead, the empirical fixation positions were binned by assigning them to the closest position every 1.85 º (7x7 grid) in the ~15x15 º portion of the face that was not covered by a mask (bottom right image of Figure 5a). The bins with a fixation frequency smaller than .02 were discarded, after which the other bin frequencies were renormalized to add up to 1. A random subset of the foveated copies made for each particular fixation position were then used proportional to the relative frequency of fixations at that position (top row of Figure 5a) compared to the fixation position with the highest frequency of fixations. The same binning and frequency normalizing procedure was also done for inverted faces (Figure 5b), using empirical human fixation positions in for inverted faces measured in humans in Experiment 1. Arguably, humans occasionally recognize faces from different orientations. For example, while lying down, the face of an upright person will present itself in an unusual orientation (rotated by 90 or 180 degrees). Similarly, recognizing somebody in an usual position both in real life and in media (such as when looking at a phone), requires processing faces at an angle that is not upright. To capture this small exposure to unusual orientations, we trained the CNN for all runs with a small proportion (0.5%) of the foveated inverted faces.

Testing.

The same testing stimuli sets were used for all runs of all models. Here, we use the word "model" to refer to the same CNN network with a specific set of parameters (learned weights between different neuron-like nodes). A difference between trained models can

occur if they are either trained with different stimulus sets, or they are retrained with the same stimulus set starting with a random set of parameters from the beginning. The testing face stimuli were different from the training stimuli used. The testing stimuli were taken from the last 30 (of 35) individual frames from a subset of the same face stimuli that were shown in movies of facial expressions to human participants in Experiment 1. This subset included the same 3 face configurations that were used as a subset for Experiment 2 (upright faces, inverted faces, and faces with a switched eyes, mouth). For each configuration 1800 individual frames (20 identities, 3 emotions per identity, 30 frames per identity and emotion) were used as a base set that was then foveated at 4 different positions. In the end, the testing stimuli consisted of 12 different stimulus sets that included the three different configurations, each of which included 4 subsets that were foveated at different positions along the vertical midline of the face, roughly corresponding to the forehead, eyes, nose, and mouth (Figure 5c).

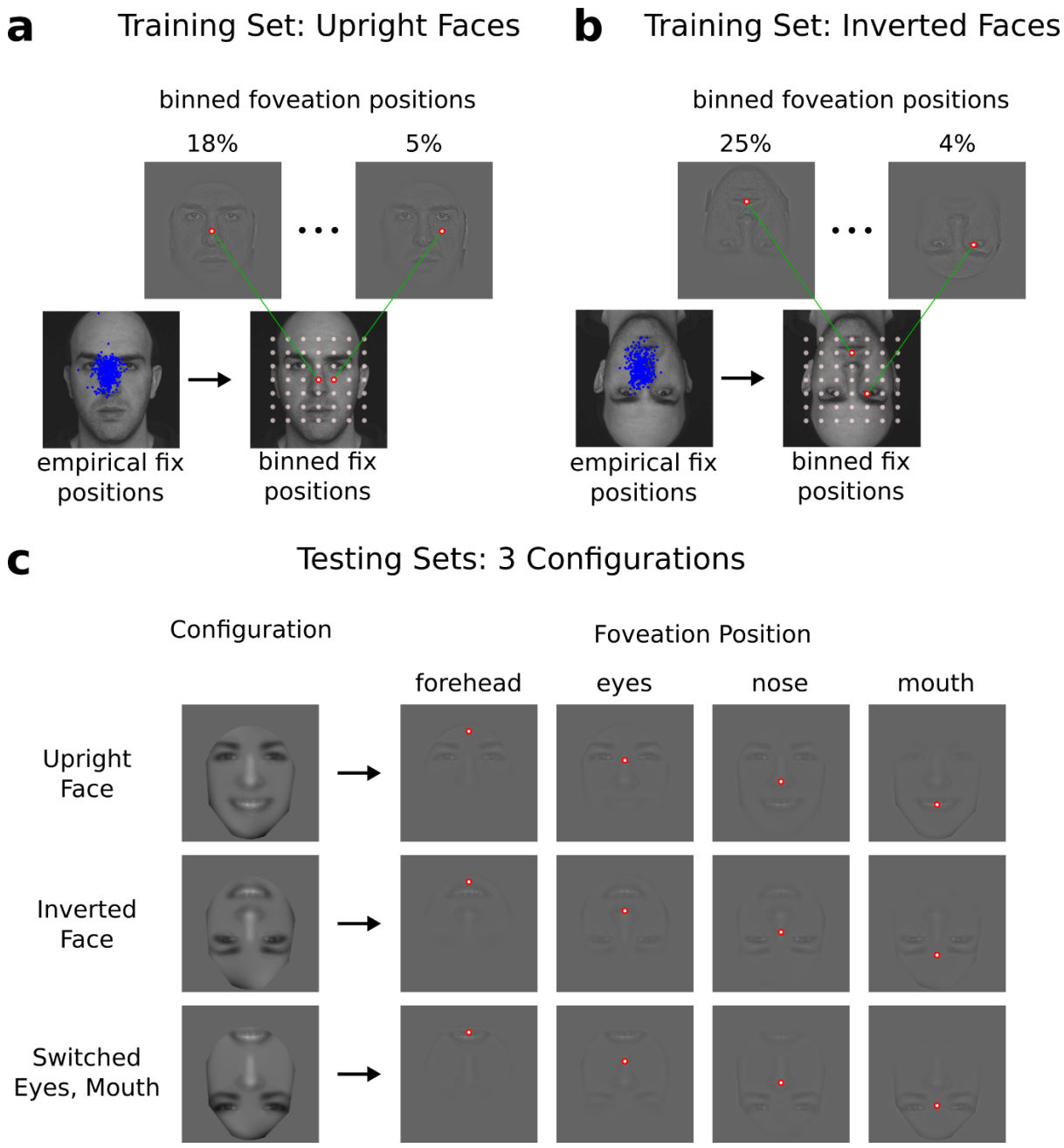

Figure 5: A summary of the simulated foveation operations done to create the training and testing stimuli sets for the CNN. ***a)*** *The bottom left image shows an example of initial fixation positions to an upright face stimulus from a single participant in the free-viewing condition of Experiment 1. Fixations from a single participant are only shown for illustration purposes. However, initial fixations from all participants were combined in order to create a training set with foveated faces that more accurately correspond to empirical fixation positions. The empirical fixation positions were binned into 49 assigned positions using a Euclidian distance measure. The grid of points in the bottom right image show the assigned positions, which cover the face. The frequency of fixations at each position was then calculated, and all positions that had a frequency of less than 2% were discarded, after which the remaining*

frequencies were normalized to sum to 1. The top left and top right images show simulated foveations centered at a frequent and infrequent fixation position, respectively, with the frequencies assigned to those positions, above each image. A training stimulus set of upright faces was then created with the number of images foveated at different fixation positions that corresponded to their empirical frequencies. ***b)*** *The same process was used to create a training stimulus set of foveated inverted faces. Small random subsets of the foveated inverted faces stimulus were then used during training of certain iterations of the CNN.* ***c)*** *After training on a foveated stimulus set with different proportions of inverted faces, each CNN model was then tested on 12 different stimulus sets that included 3 different configurations (an upright face, inverted face, and a face with switched eyes, mouth), each of which included 4 subsets that were foveated at different positions along the vertical midline of the face, roughly corresponding to the forehead, eyes, nose, and mouth.*

Architecture and Settings.

Original resnet-18.

We use an 18-layer resnet-18 (K. He, Zhang, Ren, & Sun, 2015) architecture to run a 3-class emotion discrimination task with single frames of faces (taken from movies). The network is made up of 4 "residual blocks," each of which contain 2 pairs (this number is higher for other variants of this network structure) of the same layer structure (same size and depth of feature maps) (Figure 6a). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem.

We use mini-batch (200 images per batch) stochastic gradient descent (SGD) along with a cross-entropy loss function to optimize the parameters in the model. We use hyperparameter settings of 5e-4 for the learning rate and .9 for momentum. Although this network can theoretically be run with any input image size, here we run it with an image size of 112x112 pixels due to a limitation of the resolution of the training images we used. Upscaling the training images would only increase computation time without improving

performance.

Estimating Featural information of the CNN via Class-Specific Activations Visualizations.

In addition to the original resnet-18 network, we use also the methodology of (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) and run a modified version of the same network in order to be able to construct a visualization of the important features in the input stimuli that are used by the network to do the emotion classification task. This visualization allows us to observe possible limitations in the ability to use the internal face representations that the CNN learns from (99.5% upright face and .05% inverted faces) when it is tested on face configurations that deviate significantly from the training set. The visualizations are found by mapping a weighted linear combination of the 14x14 feature maps of the last convolutional layer of the network onto the original 112x112 input images. The weights used to combine the feature maps come from the learned connections between the Global Average Pooling (GAP) layer, which acts as a unidimensional representation of the 14x14 feature maps preceding it, and the class scores output by the network. For each of the three classes, a specific Class Activation Map (CAM) is found by using the weights connecting the GAP layer to a specific class. Although (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) used this method for localization of objects in complex classification tasks with a large number of classes, it is still useful for our purpose of visualizing the features of faces that are most discriminative for the network during this task. Since the faces are aligned during both the training and testing phases, the discriminative features should be located in specific areas across CAMs. We average the visualizations across CAMs to get a single visualization map for each testing set to get an overall representations of which face features the network is able

to use the most during this task.

Figure 6b shows the modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the third and fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering the stride from 2 to 1 during the convolution operation before the last 2 residual blocks. The difference in the modified network is outlined in red.

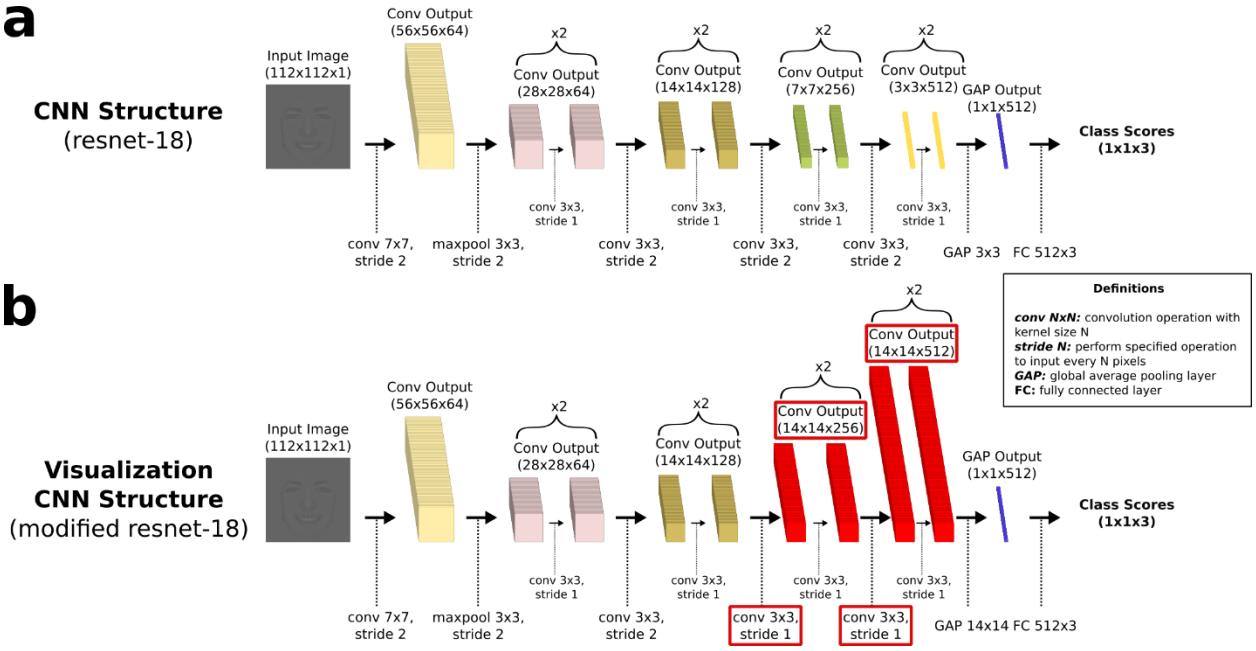

Figure 6: A flowchart of the operations and outputs involved in the CNN network that we use. **a)** *The top flowchart shows the structure and operations involved in the original resnet-18 network* (K. He et al., 2015). *Although this network can theoretically be run with any input image size, here we run it with an image size of 112x112 pixels and show the sizes of feature map outputs after max pooling and convolution operations along with the chosen depths of the feature maps at each layer, which are fixed parameter settings. Similarly, although the network is able to learn to classify an arbitrary number of classes, here we show an output of class scores for a 3-class emotion discrimination ask. One aspect of the resnet network that isn't explicitly shown in the flowchart is the "skip-connections" between layers of the same size. The network is made up of 4 "residual blocks," each of which contain 2 pairs of the same layer structure (same size and depth of feature maps). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem.* ***b)*** *The bottom flowchart shows a modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the third and*

fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering the stride from 2 to 1 during the convolution operation before the last 2 residual blocks. The difference in the modified network is outlined in red. We implement this modification in order to output a set of 14x14 pixel feature maps instead of 4x4 pixel feature maps. This allows us to use the methodology of (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) *to construct a visualization of the important features in the input stimuli that are used by the network to do the emotion classification task. This is done by mapping a linear combination of the 14x14 feature maps onto the original 112x112 input images.*

2.3 Results

Performance variations across face configurations cannot be solely predicted by the influences of feature proximity to an optimal point of fixation

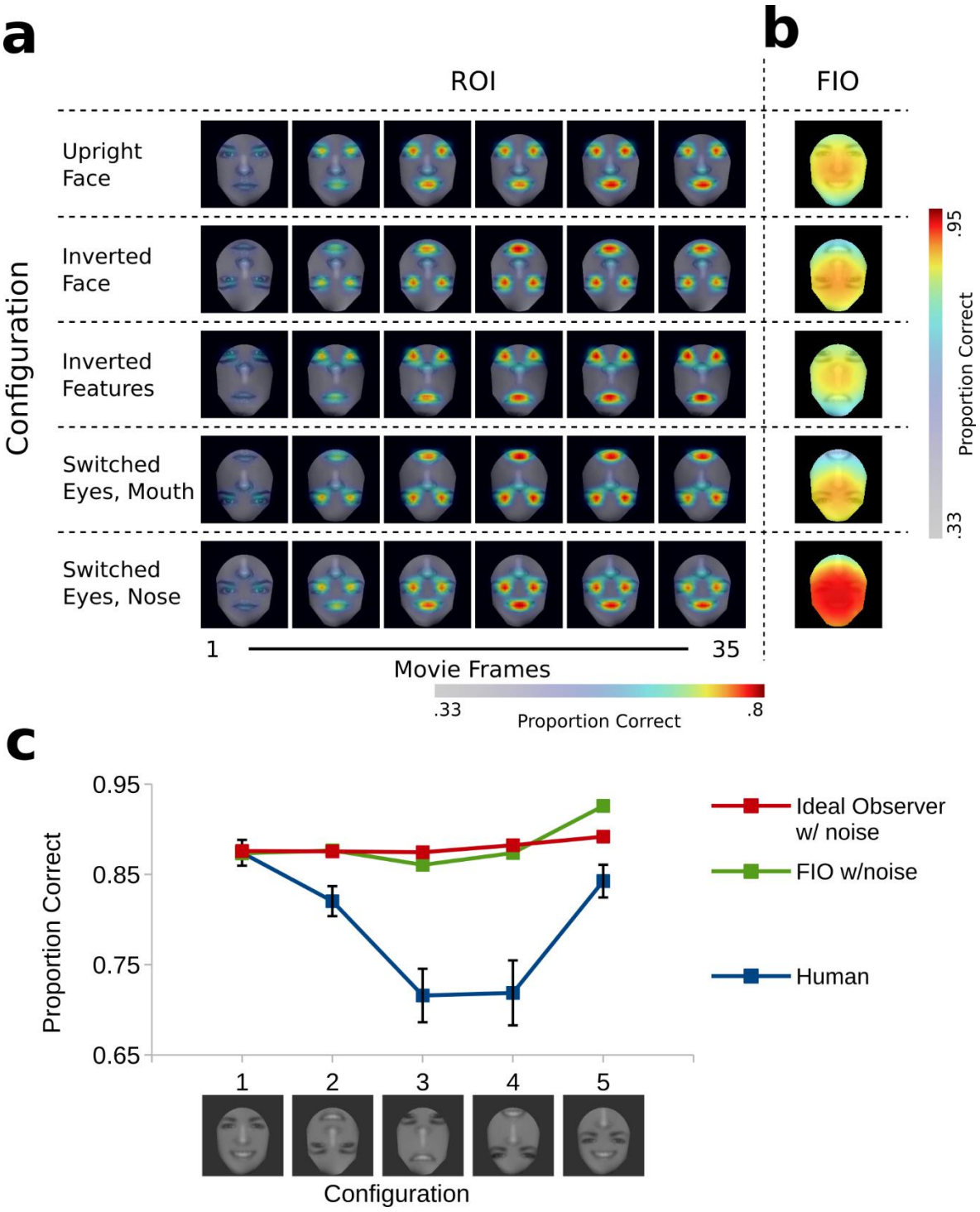

Figure 7: ***a)*** *The results of the ROI analysis are shown for every 6th frame of each configuration. The ROI analysis was done for each frame separately (i.e, the task was simulated with only frame n from each of 60 face expression movies and then repeated for all n). For all configurations, most of the information for the task is contained in the eyes and mouth. The information content for the emotion discrimination task increases across roughly the first third of each movie and then remains stable for the rest of the movie. The only difference between the configurations is the location of the informative features, rather than*

changes in the overall information content of the faces. ***b)*** *The results of the FIO model for the first fixation are shown for each configuration. Despite the mouth being an important feature, the theoretical optimal initial fixation position for different configurations is located near the eyes, even for those configurations where the mouth is located far from the eyes, because they are two important features.* ***c)*** *Performance (proportion correct) in the emotion discrimination task with 5 different face configurations is shown for the free-viewing condition in Experiment 2 in blue, the theoretical optimal points of the FIO in green, and for an ideal observer in red. Noise was added to both the ideal observer and the FIO in order to fit the human performance for the upright faces and compare how performance changes across the different configurations. The stimuli were manipulated such that the overall information content for each configuration would be roughly equal, which is reflected in the steady performance of the ideal observer. The FIO, however, shows some changes in performance, especially an increase at the fifth configuration, where there are two important features (eyes and mouth) for this task located close together. However, human performance shifts much more considerably between configurations.*

In Experiment 1, we ran 11 observers in an emotion discrimination task with a free eye movement condition with 5 different blocked face configurations (Figure 2a) and compared their preferred fixation positions to theoretical optimal fixation positions predicted by the FIO model. First, in order to determine which features are important for the emotion discrimination task and how their importance changes over time in all 5 of the face feature configurations, we run an ROI model frame by frame for each configuration. The ROI model outputs a performance map that shows how an ideal observer performs using only small windows at a time centered at different parts of the stimulus. Figure 7a shows the results of the ROI analysis for every 6th frame of each configuration during the 1400ms (35 frame) presentation period in the free eye movement condition. For all configurations, most of the information for the task is contained in the eyes and mouth. The information content for the emotion discrimination task increases across roughly the first third of each movie and then remains stable for the rest of the movie. The only difference between the configurations is the location of the informative features, rather than changes in the overall information content of the faces. We then run an FIO model and output a performance map for possible fixation

positions to the stimulus, for each configuration, as shown in Figure 7b. The FIO model predicts a change in the theoretical optimal point of initial fixation in face configurations that depends on the positions of the important features for the task (eyes and mouth). When the eyes and mouth are positioned relatively close together, as in the upright face configuration, the inverted face configuration, and especially the configuration with a switched position of the eyes and nose, the theoretical optimal point is located between them but a bit closer to the eyes. However, when the eyes and mouth are positioned far apart from each other, the theoretical optimal point is much closer to the eyes because even though the mouth contains a large amount of information, there is more information contained in the combination of the two eyes together.

Humans show large differences in performance between different face configurations, as shown in Figure 7c. Figure 7c also shows performance in the task for a standard Ideal Observer model, as well as for an FIO at the theoretical optimal points. For both the ideal observer and the FIO, the level of contrast used to run the stimuli is matched to the level used in the human experiments. However, a large amount of noise is added to the Ideal Observer, and a smaller amount of noise added to the FIO in order to match performance in both models with human performance for upright face configuration. This is done in order to use it as a reference point when comparing performance differences across configurations in humans vs performance differences across configurations in each of the models. The face stimuli were manipulated such that the overall information content for each configuration would be roughly equal, which is reflected in the steady performance of the ideal observer. The FIO, however, shows some changes in performance between configurations, especially an increase at the fifth configuration, where there are two important features (eyes and

mouth) for this task located close together. However, human performance shifts much more considerably between configurations. In addition, even though for the fifth configuration, where an FIO shows an increase in performance, human performance is still significantly lower compared to the upright face configuration, $t(10) = 3.79$, $p = 4.3E\text{-}3$, one-tailed. This suggests that the interaction of the foveated visual system with the location of important features for the task, does not explain much of the differences in performance seen when humans are presented with faces that have shifted features.

Performance variations across face configurations cannot be explained solely by inability to learn theoretical optimal eye movements.

The analysis above assumes that human eye movements are programmed to the theoretical optimal location for each configuration at which the most information can be extracted. It is known that humans, on average, direct their first eye movement to the theoretically optimal spot for intact, upright faces. However, there is a possibility that humans might not learn to make saccades to the new optimal locations that are predicted by the FIO for other face configurations. This may lead to further degradation of performance with unusual face configurations. As seen in Figure 8a, although the FIO model predicts a change in the position of the theoretical optimal point of fixation, most of the human participants do not fixate this point in at least two of the four face configurations, besides the upright one, that were run in Experiment 1. One possibility is that due to the face mask that we use (to cover the external face features like ears and hair), observers simply keep their preferred initial point of fixation in other configurations the same as in the upright face configuration. However, there is a statistically significant effect that shows an interaction of face configuration and preferred vertical fixation position, $F(4) = 7.21$, $p = 9.67E\text{-}5$.

To take into account the suboptimal initial fixation strategies of humans, we implemented a version of the FIO, a Fixation-Weighted FIO (FW-FIO), that models eye movements to each configuration based on the empirical distribution of human initial fixations. Figure 8b shows a flowchart for the procedure of calculating performance with the FW-FIO. Figure 8c shows that the FW-FIO is able to model larger differences in performance between face configurations relative to the FIO, but those differences are still much smaller relative to human performance.

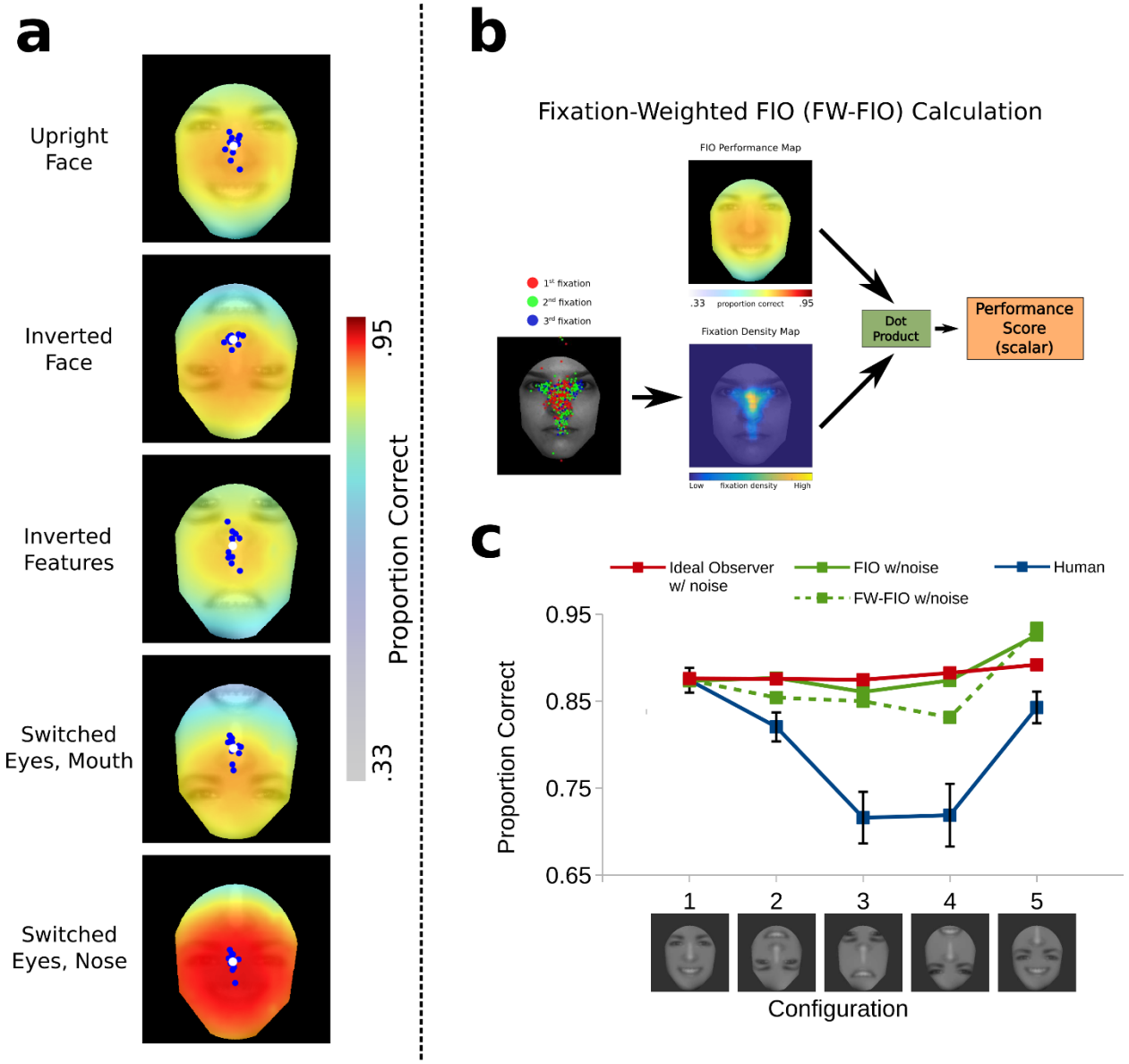

Figure 8: ***a)*** *The results of the FIO model for the first fixation are shown for each configuration, with averages of empirical initial fixation positions from observers overlaid in blue, as well as an average across observers shown as a white point. Despite the mouth being an important feature, the theoretical optimal initial fixation position for different configurations is located near the eyes, even for those configurations where the mouth is located far from the eyes, because they are two important features. Observers' empirical initial fixation positions differ from the theoretical optimal initial fixation positions to varying degrees in different configurations.* ***b)*** *A flowchart is shown for how a performance score for the FW-FIO models is calculated for each face configuration. Here, an intact, upright face configuration is shown as an example. An empirical fixation distribution is compiled from all observers for a specific face configuration. A fixation density map is then calculated by binning fixation positions into discrete locations on the face image and normalizing such that the fixation frequencies across the face image sum to 1. Here, a*

smooth fixation density map with every possible fixation location is shown for illustration purposes. An FIO performance map is then used such that a weighted sum is taken from different locations of the performance maps to output a single scalar value. The locations of the performance maps correspond to the locations of the bins that were used to bin the empirical fixation positions. The weights at those locations correspond to the normalized fixation frequency values. ***c)*** *Performance (proportion correct) in the emotion discrimination task with 5 different face configurations is shown for the free-viewing condition in Experiment 2 in blue, the theoretical optimal points of the FIO in solid green, for an FW-FIO in dashed green, and for an ideal observer in red. Noise was added to the ideal observer, the FIO, and the FW-FIO in order to fit the human performance for the upright faces and compare how performance changes across the different configurations. The stimuli were manipulated such that the overall information content for each configuration would be roughly equal, which is reflected in the steady performance of the ideal observer. The FIO, however, shows some changes in performance, especially an increase at the fifth configuration, where there are two important features (eyes and mouth) for this task located close together. The FW-FIO, which takes human empirical fixation positions into account shows larger differences in performance relative to the FIO. However, human performance shifts much more considerably between configurations compared to all of the models.*

Evaluating configural representations through comparisons against the FIO.

Our analysis of the preferred initial fixation positions of humans suggests that they do not seem to adapt their initial eye movement for most of the unusual face configurations to the theoretically optimal fixation points found with the FIO. However, another possibility is that the empirically optimal eye movements for humans do not correspond to those predicted by the FIO. The FIO assumes a perfect representation of face features, regardless their spatial configuration in a face template, as long as the distance of those features to a specific fixation position remains the same. In contrast, humans might have additional constraints and inefficiencies in how they extract information from different features in different configurations. If this is the case, then we would expect that human performance as a function of fixation position might depart from that predicted by the FIO. Furthermore, the human empirical optimal point of fixation might not be predicted by the FIO. Below, we describe results of measuring human performance as a function of fixation position and comparing it to the FIO performance profile for different configurations.

Intact, Upright Face Stimulus.

In Experiment 2, we run 7 observers on both a free eye movement and a forced-fixation condition with a subset of 3 of the 5 configurations used in Experiment 1. The configurations used are an intact upright face, an inverted face, and a face with a switched position of the eyes and mouth. We measure human performance when forced to fixate at 4 positions, corresponding to the forehead, eyes, nose, and mouth, down the vertical midline of the face for each configuration. We then compare the human performance profile to the performance of an FIO. As seen in Figure 9a, for intact upright faces there is an area of

relatively flat performance in the lower part of the face across the eyes, nose, and mouth ($t(6) = 2.08$, $p=.08$, one-tailed for eyes vs. nose; $t(6) = 1.37$, $p=.22$, one-tailed for nose vs. mouth) and significant drop in performance at the forehead ($t(6) = 5.89$, $p=1E-3$, one-tailed for eyes vs. forehead). This performance profile is more flat (has smaller differences in performance between points) but is consistent with an empirical optimal fixation position around the tip of the nose previously found in a 1 of 7 emotion discrimination task with static images of faces (Peterson & Eckstein, 2012). In addition, participants fixate close to this point on average, as seen with the blue bar in Figure 9a. Since the performance profile is more flat, rather than using FIO parameters that were originally fit to a face identification task (Peterson & Eckstein, 2012), we fit the parameters of the FIO to the current emotion discrimination task with dynamic faces in the upright intact face configuration. We then use the same parameters to run the FIO model with the inverted face configuration and the face configuration with a switched eyes and mouth to see if the model is able to predict human empirical optimal fixation positions.

Inverted Face Stimulus.

For clarity, in the description of feature locations for face configurations that aren't upright, we will refer to features that have been moved relative to an upright face as "new." As seen in Figure 9a, for inverted faces, there is better performance at the top of the stimulus, which corresponds to the new tip of the nose and new mouth on an inverted face stimulus ($t(6) = 2.96$, $p=.025$, one-tailed for nose vs. eyes; $t(6) = 6.43$, $p = 6.67E-4$, one-tailed for eyes vs. forehead) (the labels in the Figure 9a are shown relative to an upright face stimulus). In addition, the overall performance level at all points is lower relative to an upright face stimulus. However, the performance profile of the FIO with an inverted face stimulus does

not exhibit this drop in performance, especially in the lower part of the stimulus, and predicts a theoretically optimal point in the new eye and new forehead region of inverted faces, which does not match the location of the empirically optimal point in the new nose and new mouth region. In addition, in the free eye movement condition, human participants fixate significantly closer to the empirically optimal region of inverted faces relative to the theoretically optimal region found with the FIO, as seen in the location of the green bar in Figure 9a. We also fit an FIO model with the same parameters except with higher noise to see if it is better able to capture the human performance curve when is in a closer performance range compared to humans, as seen in Figure 9b. However, although the shape of the new FIO performance curve for inverted faces is a bit flatter, it still predicts the same region for the theoretically optimal point compared to the model with a lower noise level, which does not represent the human performance profile well.

Switched Eyes, Mouth Face Stimulus.

As seen in Figure 9a, for the third configuration (switched eyes and mouth) performance is relatively flat across all points, with a small decrease at the nose only relative to the new mouth ($t(6) = 2.68$, $p = .037$, one-tailed for nose vs. new mouth; $t(6) = 1.29$, $p = .24$, one-tailed for nose vs. new eyes). In addition, the overall performance level at all points is lower relative to both an upright face stimulus and an inverted face stimulus. However, the performance profile of the FIO with a switched eyes, mouth face stimulus only exhibits a drop in performance at the top of the stimulus, which is still significantly smaller than what is seen in humans, and predicts a theoretically optimal point in the nose and new eyes region, which does not match the flat performance profile of humans. In the free eye movement condition with a switched eyes, mouth face stimulus, human participants fixate

significantly far from the theoretically optimal region found with the FIO, as seen in the location of the red bar in Figure 9a. Similar to what is done with the FIO with inverted faces, we also fit an FIO model with the same parameters except with higher noise to see if it is better able to capture the human performance curve when is in a closer performance range compared to humans, as seen in Figure 9b. However, although the shape of the new FIO performance curve for a face stimulus with switches eyes, mouth is a bit flatter, it still predicts the same region for the theoretically optimal point compared to the model with a lower noise level, which does not represent the human performance profile well.

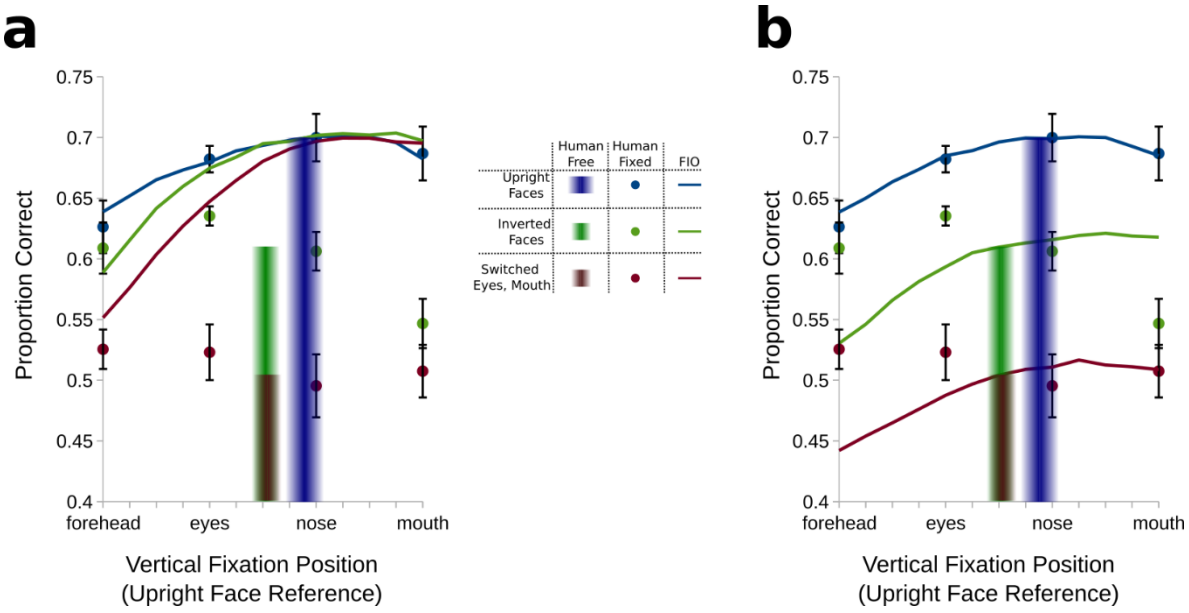

*Figure 9: (**a**) Human performance down the vertical midline of the face in the forced-fixation condition of Experiment 2 is shown for the 3 configurations that were tested, with blue points for upright faces, green points for inverted faces, and red points for faces with a switched eyes, mouth. The x-axis labels represent fixation positions relative to an upright face, so the performance for the other two configurations at those points does not actually correspond to the labels, but instead corresponds to relative locations down the vertical midline of the images. The preferred fixation positions from the free-viewing condition of Experiment 2, along with a standard error of the mean on each side are shown in corresponding colors. Performance of the FIO at fixation positions down the vertical midline of the face, run with the same noise level for all configurations, is also shown in corresponding colors. This noise level was fit to the human data for the upright configuration. The FIO does not predict the sharp drop in performance that is seen when humans do the emotion discrimination task with face configurations that are not upright. In addition, the theoretical optimal fixation position for inverted faces (in green) does not predict the empirical optimal fixation position for*

*humans. For this configuration, the human preferred point of fixation is much closer to the empirical optimal point than the theoretical optimal point predicted by the FIO. The theoretical optimal fixation position found by the FIO for the face configuration with a switched eyes and mouth does not predict the lack of an empirical optimal fixation position (differences in performance across fixation positions are not statistically significant for this configuration) in humans. (**b**) The same data is shown as in the previous plot, except that the noise level of the FIO is now fit separately to the human data for each individual configuration. Although the FIO is now fit to the overall performance level for each configuration, it is still unable to predict it is still not able to predict the location of the empirically optimal fixation positions in humans for face configurations that are not upright.*

Capturing the learning of configural representations using a Convolutional Neural Network (CNN).

As seen above, the FIO is does not provide a strong fit to the forced-fixation human data for inverted faces or for faces with a switched position of the eyes and mouth. An FIO is only limited by a simulation of a foveation visual system, but otherwise has a perfect internal representation of the faces of each configuration that it is run on. Humans, however, are overwhelmingly exposed to intact upright faces and likely have an internal representation of faces that is tuned to the kinds of face stimuli that they are exposed to. Face stimuli that are not intact and upright may be difficult for humans to process efficiently because of a mismatch between the stimuli being shown compared to an internal representation of upright face stimuli stored in the human brain. This is an important limitation that the FIO does not capture. To stimulate this limitation, we use a CNN model where we train it on 99.5% intact, upright faces, and on 0.5% inverted faces. The training stimuli contain copies of the same faces that are processed differently based on a foveation simulation at different fixation locations. Those fixation locations are taken from human empirical fixation distributions to upright and inverted faces during the free-fixation conditions (see Methods for details). Figure 10a shows the results of a CNN model (resnet-18) for a human emotion discrimination task that is tested on stimuli from each of the three face configurations used in the human forced-fixation condition. Each configuration is then separately "foveated" and tested at the same 4 locations that humans were tested on in the forced-fixation condition. Performance is highest overall for tested upright faces, then inverted faces, and finally faces with a switched eyes and mouth. Both the performance order across the 3 testing sets, and the relative order of performance at different fixation points, within each testing set, is a much

better representation of human differences in forced-fixation performance relative to the predictions of the FIO. In Figure 10b, the same type of plot is shown as in (a), except for a model, whose network architecture has been modified from a regular resnet-18, to one where the last several convolutional layers have larger dimensions of feature maps. This was done in order to be able to use the enlarged features maps from the last layer of the modified model to visualize which parts of a face stimulus the network used the most during the classification task. Enlarging the feature maps of the last several convolutional layers relative to the original CNN network results in differences in the performance profile for each configuration. In Figure 10c, a visualization is shown of the parts of the face stimuli that the CNN model trained on a face emotion discrimination task uses the most. The 3 rows represent three different face configurations that were used during testing: upright faces, inverted faces, and faces with a switched eyes and mouth, respectively. The 4 columns represent different simulated fixation positions at the forehead, eyes, nose, and mouth, respectively. The scale used to show the importance of different face features is relative only within each image because each visualization has been normalized such that the features with the highest weights are mapped to the highest values. The visualizations show that the mouth is an important region that the CNN uses to do the classification task. However, the extent to which the mouth is used by the network, depends on both the configuration being used during testing, as well as the fixation position at which at which a foveation simulation is applied.

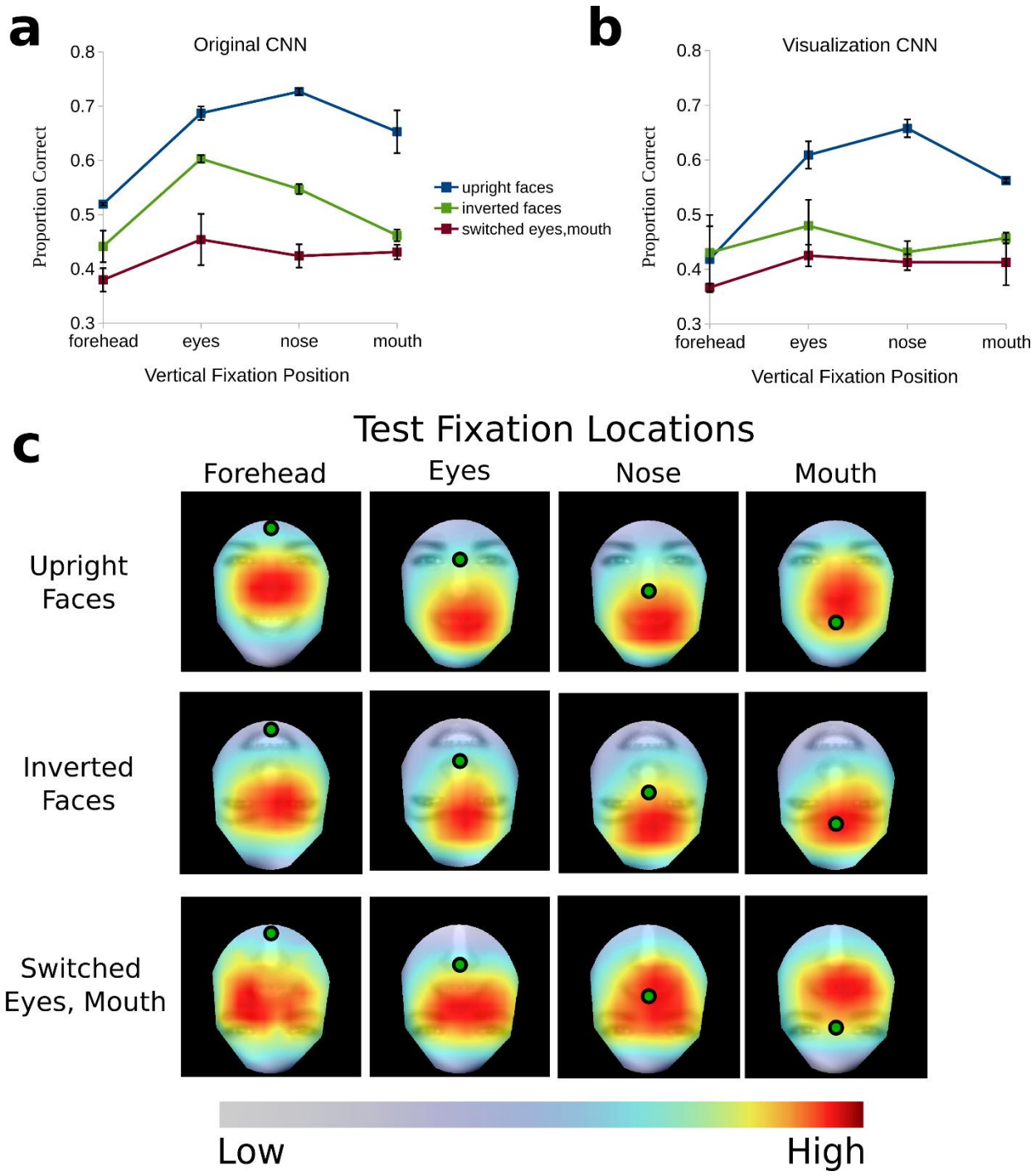

Figure 10: ***a****) The results of a CNN model are shown for a human emotion discrimination task. During the testing phase, the same sets of stimuli are used as the ones used to run the FIO model and to run the human forced-fixation psychophysics experiments, except they are separated into individual frames and are run as images instead of movies. The differently colored lines represent performance profiles for a renset-18 CNN model, separately tested with different stimulus sets, corresponding to upright faces, inverted faces, and faces with a switched eyes and mouth. Each of the testing stimuli sets were also foveated at 4 different*

positions, where the center of foveation was at the forehead, eyes, nose, and mouth, shown on the x-axis. The error bars represent standard error of the mean across 5 different training runs (i.e. a network trained from a random weight parameter setting 5 separate times). Performance is highest overall for upright faces, then inverted faces, and finally faces with a switched eyes and mouth. Both the performance order across the 3 testing sets, and the relative order of performance at different fixation points, within each testing set, is a much better representation of human differences in forced-fixation performance relative to the predictions of the FIO. ***b****) The same type of plot is shown as in (a), except for a model, whose network architecture has been modified from a regular resnet-18, to one where the last several convolutional layers have larger dimensions of feature maps. This was done in order to be able to use the enlarged features maps from the last layer of the modified model to visualize which parts of a face stimulus the network used the most during the classification task. Enlarging the feature maps of the last several convolutional layers relative to the original CNN network results in differences in the performance profile for each configuration.* ***c****) A visualization is shown of the parts of the face stimuli that the CNN model trained on a face emotion discrimination task uses the most. The 3 rows represent three different face configurations that were used during testing: upright faces, inverted faces, and faces with a switched eyes and mouth, respectively. The 4 columns represent different simulated fixation positions at the forehead, eyes, nose, and mouth, respectively. The scale used to show the importance of different face features is relative only within each image because each visualization has been normalized such that the features with the highest weights are mapped to the highest values. The visualizations show that the mouth is an important region that the CNN uses to do the classification task. However, the extent to which the mouth is used by the network, depends on both the configuration being used during testing, as well as the fixation position at which at which a foveation simulation is applied.*

2.4 Discussion

Previous studies have shown that there is a decrease in performance during various face discrimination tasks when face features have been shifted, rotated, or altered in some way relative to an intact upright face (Farah et al., 1995, 1998; Tanaka & Farah, 1993; R. K. Yin, 1969; Young et al., 1987). This performance difference has generally been attributed to a disruption of the use a face template represented in higher-order face-specific brain areas (Farah et al., 1998; Tsao & Livingstone, 2008), but see (Richler, Cheung, & Gauthier, 2011; Richler et al., 2012) for an alternative explanation based on a disruption of a practiced attentional strategy. However, few previous studies of holistic face processing (Bombari et al., 2009; Heering et al., 2008; Hills et al., 2013; Van Belle et al., 2010; Xu & Tanaka, 2013) have controlled for an important low-level aspect of the visual system that is known to affect performance in upright faces. This aspect is a foveated visual system, which has been shown to guide the initial eye movement during various face discrimination tasks to a specific optimal (performance-maximizing) location on a face. In previous studies with intact, upright face stimuli, performance has been shown to decrease in upright faces when observers were forced to fixate locations other than their empirical optimal spot. Here, we asked whether the performance differences previously found between the use of intact upright face stimuli and altered face stimuli may be partly explained by an altered interaction of a foveated visual system with the altered face features. We reproduced large differences in performance in a face discrimination task between an upright face stimulus and four other face stimuli with altered features. However, we found that very little of these differences in performance could be explained purely by disruptions in foveated processing.

We found that when humans freely fixate different face configurations, they are unable to take advantage of the same information contained in important face features for this

task, when the features are shifted, but remain the same or even a shorter relative distance away from each other. This contrasts with ideal observers and foveated ideal observers, which are only affected by the overall information content in an image, or different distances of important features to each other, respectively. The most informative example of this was the result obtained with the fifth configuration, where the most important features for this task, the eyes and mouth were located close to each other. The FIO showed a significant increase in performance at the theoretical optimal spot for this configuration relative to the intact upright face. Human participants, however, showed a decrease in performance despite fixating the theoretical optimal spot in the free eye movement condition. It should be noted though that human performance in this configuration was significantly higher than in the other altered configurations. This suggests that the relative locations of informative features to each other may be an important consideration when interpreting the cause of the magnitude of performance differences between different face configurations. However, the fact that even with a location advantage of the features, human performance still suffered, points to a significant contribution of a disruption in higher-order face processing mechanisms.

We assessed another possible contribution of foveated visual processing to performance differences in different face configurations: suboptimal initial fixation positions. We found with a large number of participants, that the initial fixation was enough to achieve asymptotic performance in this task with upright faces. Previous studies have also shown the initial fixation position in upright faces is made to a consistent location that is similar across people and is resistant to change when it becomes suboptimal because of a change in the type of face stimuli that are shown or a change in the visual system. In relation to this, we asked

if humans make suboptimal initial fixation positions to face configurations that they don't have prior experience with. However, suboptimality can be defined relative to a theoretical optimal position found with an FIO, or relative to an empirical optimal position found in humans when forced to fixate different locations on the face. Previous studies have shown that for upright faces, the empirical optimal initial fixation position is predicted well by a theoretical optimal position found with the FIO and observers fixate this location in a free eye movement condition. In contrast, we found that in three of the four altered face configurations, observers did not fixate the theoretical optimal spot. Using further analysis in a forced-fixation condition with two of those three configurations, we found that the FIO did not correctly predict the location of an empirical optimal fixation position. In one of those configurations (switched position of eyes and mouth), an empirical optimal point did not exist at all, since performance was relatively flat at different fixation positions down the vertical midline of the face. However, in the other tested configuration (inverted face), an empirical optimal fixation position did exist and observers fixated close to it when they were allowed to freely fixate. These results again suggest that suboptimal fixation positions may only play a minor role in performance differences between different configurations. The fact that the FIO is unable to predict empirical optimal fixation positions for human observers points to a decreased efficiency of integrating information from features when they are not located in their expected positions in an upright face. In addition, the fact that human observers fixate close to their empirically optimal spot in the inverted-face configuration, a commonly used one in studying holistic processing, shows that even detriments relative to their own optimal performance are only a minor factor in the overall performance differences observed between different configurations.

In order to help explain the failure of the FIO to predict human performance differences between different face configurations, we use a CNN model to simulate the higher-level aspects of the visual system. Although they are only a rudimentary approximation of human cortical processing, CNNs are starting to be used in the study of human vision and face processing (see (O'Toole, Castillo, Parde, Hill, & Chellappa, 2018) for a review) after successful implementations of various face classification tasks in computer vision (Li, Lin, Shen, Brandt, & Hua, 2015; Schroff, Kalenichenko, & Philbin, 2015; Taigman, Yang, Ranzato, & Wolf, 2014) , some of which have achieved close to human performance. CNNs are known to have certain useful properties that may be able to represent aspects of the human visual system. One of those aspects is a feedforward multilayer structure that represents progressively more complex features starting from edge detection and ending with complex shapes, textures, colors, and the relationships between them. Another important aspect is the ability to learn feature detectors that are adapted to the complex statistical properties of the features in the images that the model is being trained on. CNNs are also able to represent objects with some degree of spatial location invariance due to pooling operations between different layers. However, a large part of classification performance on new test stimuli depends on their similarity in multiple feature dimensions, including feature location, to the training stimuli that the CNN was exposed to. In the context of face processing, humans are overwhelmingly exposed to upright intact faces that are foveated through retinal processing before more complex features are extracted further in the visual stream and eventually in cortical areas. In order for the CNN model to represent faces as accurately as possible relative to human experiences with them, we use empirical fixation data with an overwhelming majority of upright faces to create foveated training stimuli. As a

result, the CNN model that we train is likely able to represent something similar to a configural representation, where the location of important features for this task is an important feature that degrades performance when it is disrupted. We show that a CNN model is much better able to predict human empirical optimal points of fixation as well as overall performance profiles (performances at different points of fixation along the midline of the face in a forced-fixation condition), compared to an FIO for the three configurations that were tested in the forced-fixation condition. When processing faces, the performance of an FIO is only limited by a simulation of foveation, because it then makes optimal decisions further up in the processing stream. In contrast, the performance of the CNN model that we use is limited both by a simulation of foveation (because of pre-foveated input images) as well as a complex representation of face features that are learned during training.

2.5 Supplementary Materials

Here we describe the methods and details of a third experiment that we ran, which informed our use of the Bayesian Ideal Observer and Foveated Ideal Observer models.

Materials and Methods.

Participants.

The third experiment was a free viewing condition with a gaze contingent stimulus presentation time and a single upright face configuration and was completed by a separate group of 91 students. Data from 26 of those students was discarded because they performed either at chance level, or below chance level by consistently misclassifying one of the 3 emotions. Informed consent was obtained from all subjects and guidelines provided by the institutional review board of the University of California, Santa Barbara were followed.

Apparatus and Materials.

The same eyetracking and monitor setup was used as for the two experiments presented in the main text.

Stimuli.

In this experiment, only a subset containing upright face configurations was used from the stimuli set in the two experiments presented in the main text.

Stimulus Presentation on Screen.

Each frame of each movie was luminance-mean normalized to 62 cd/m² and shown to participants at a Root Mean Square (RMS) contrast of .1441, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 6.83 cd/m² (corresponding to a noise RMS contrast of .11). Participants viewed the face stimuli 76.5cm

away from the display resulting in a square stimulus (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. The large size of the faces, more typical of conversational distance, was chosen: (1) to allow measurements of larger variations of perceptual performance with fixation position (for small faces perceptual performance is less sensitive to fixation position within the face); (2) to allow more precise measurements of fixation positions relative to facial features. In addition, the large faces (10 deg. width, 15 deg. height) have been shown to be the face size that optimizes face identification (Yang et al., 2014).

Procedure.

Observers performed a gaze-contingent free-eye movement emotion identification task with 3 total blocks consisting of a single upright face configuration. Each block consisted of 125 trials. The contrast (described in the *Stimulus Presentation on Screen* section above) remained the same for all the blocks. However the stimulus presentation time varied randomly from trial to trial, with stimulus presentation ending after the start of a second, third, or fourth saccade with equal probability. All trials ended after a maximum presentation time of 1400ms if a participant did not have time to start the saccade number that was randomly chosen for that trial.

Experimental Conditions.

Starting Fixation.

During free-viewing blocks, participants started a trial by pressing the space bar while fixating a cross (.44 º x .44 º) in one of eight randomly chosen locations located on average 13.94º from the center of the stimulus. The fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants

moved their eyes more than 1° from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus.

Trial Timing.

The stimulus was presented after an initial fixation cross. The stimulus presentation time varied randomly from trial to trial in the free eye movement condition of this experiment, with stimulus presentation ending after the start of a second, third, or fourth saccade with equal probability. All trials ended after a maximum presentation time of 1400ms if a participant did not have time to start the saccade number that was randomly chosen for that trial. At the end of each trial, participants had unlimited time to select with the mouse one of three emotions that were displayed on the screen. No feedback was given.

Results.

Performance saturates after a single eye movement in the dynamic emotion discrimination task.

For all experiments presented in this paper, we used a dynamic stimulus set (movies of faces) to realistically represent the temporal dynamics of the unfolding of a facial expression. In Experiment 3 we use a gaze-contingent free eye movement paradigm where the stimulus presentation is randomly interrupted after either one, two, or three eye movements. The experimental manipulation allowed to assess the role of increasing number of eye movements and presentation times on perceptual performance. For this experiment we only used an upright face configuration as the stimulus set. Figure S.1 shows that there is not a statistically significant improvement in performance after the first eye movement, $t(64) = 1.3902$, $p = 0.085$, one-tailed for performance after one eye movement vs. after two eye

movements, and t(64) = 0.917, p = 0.181 for performance after one eye movement vs. after three eye movements. The result suggests that the majority of the visual information supporting the perceptual task is extracted during the first fixation.

This result informs the use of a short presentation time and the representation of a single eye movement when running the FIO model in order to more accurately represent the use of only a small time frame to extract most of the information for the task. Although we use a simulated short presentation time (only 5 frames) for the FIO model, we still use a longer presentation time (20-35 frames) in the free eye movement conditions of Experiment 1 and Experiment 2, in order to provide enough time for participants to make an initial eye movement as well as to provide a more ecologically valid facial expression viewing time when making free eye movements.

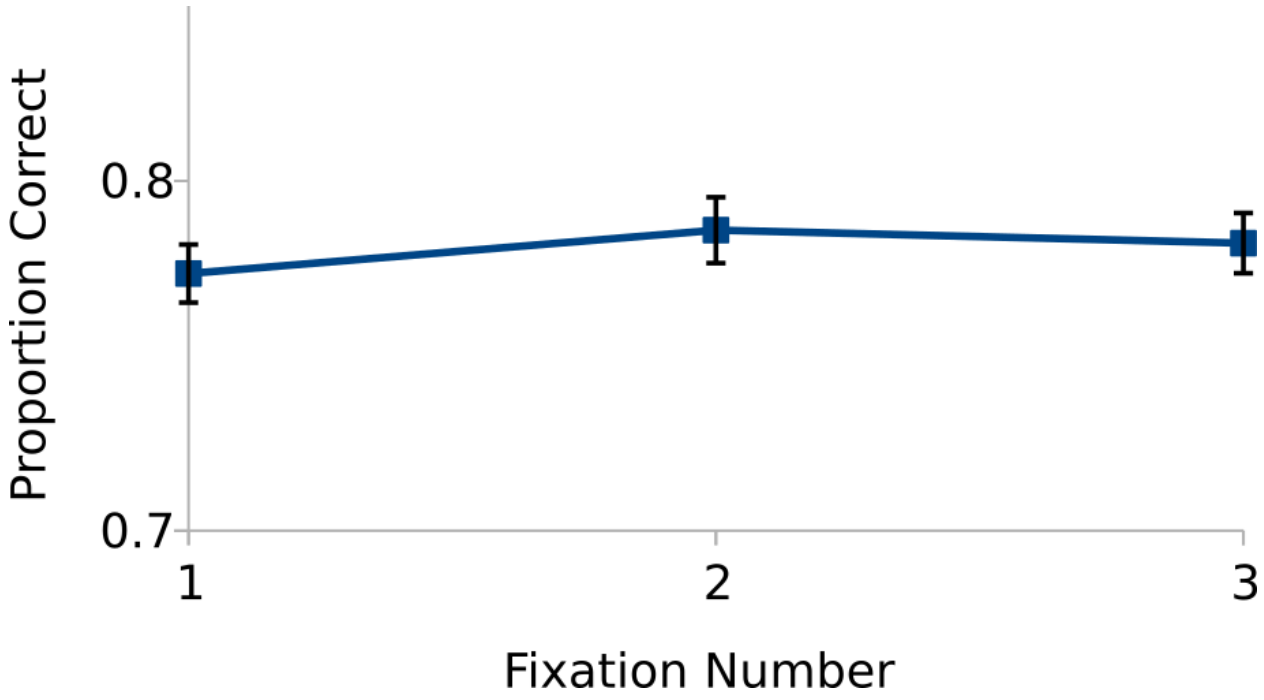

Figure S.1: The results from Experiment 3 are shown. Performance in the emotion

discrimination task is plotted as a function of the number of fixations that were allowed in a trial. Differences in performance between trials where one, two, or three saccades were made are not statistically significant, showing that the vast majority of information for this task is gathered before the second fixation.

3 Eye movements during gender discrimination of faces are adapted to the naturally occurring statistics of emotional expressions

3.1 Introduction

Face perception is a ubiquitous task that most humans perform many times a day. Eye movements, which point the high acuity foveola during exploration of the environment, are critical for accomplishing this evolutionary important task. For a variety of visual tasks, ranging from visual search to reading, the human brain programs eye movements by taking into account the foveated properties of the visual system in conjunction with the distribution of task-relevant information in the environment (M. P. Eckstein, Schoonveld, Zhang, Mack, & Akbas, 2015a; G. E. Legge, Klitz, & Tjan, 1997; Gordon E. Legge, Hooven, Klitz, Stephen Mansfield, & Tjan, 2002; Najemnik & Geisler, 2005, 2009; Paulun, Schütz, Michel, Geisler, & Gegenfurtner, 2015; M. F. Peterson & Eckstein, 2012) to maximize the acquisition of information during basic perceptual tasks (optimal or near-optimal oculomotor strategies; but see (Morvan & Maloney, 2012; Verghese, 2012) for suboptimal strategies).

During face discrimination tasks, humans exhibit high levels of accuracy with even a single saccade (Hsiao & Cottrell, 2008; Or, Peterson, & Eckstein, 2015). The majority of

humans direct their first fixation to a featureless point, just below the eyes, that maximizes accuracy in evolutionarily important perceptual tasks such as face identification, gender discrimination, and emotion discrimination (M. F. Peterson & Eckstein, 2012). There is also evidence for increased neural activity in the posterior lateral face patch in monkeys (Issa & DiCarlo, 2012) and increased separation of activity patterns to major face features in the right inferior occipital gyrus in humans (de Haas et al., 2016), when the facial features appear at the typical retinal positions relative to the preferred point of fixation compared to atypical locations.

Furthermore, the preferred point of initial fixation to a face varies moderately across tasks and is predicted by a theoretical model (the Foveated Bayesian Ideal Observer; FIO) that takes into account the relevant information in the faces for a given perceptual task as well as the foveated characteristics of the visual system and integrates information across the visual field. For example, for an emotion identification task, which contains more information in the mouth area, the human initial fixation to the face is directed to a lower point along the face than for an identification task. In addition to differences in the initial fixation position between tasks, there are also individual differences across observers that correspond to observer-specific optimal fixation positions (M. F. Peterson & Eckstein, 2013) that are consistent across time. Beyond the initial eye movement, scanpaths involving multiple eye movements are also known to be idiosyncratic both for individual participants and specific tasks, and are consistent across time (Kanan, Bseiso, Ray, Hsiao, & Cottrell, 2015; Mehoudar, Arizpe, Baker, & Yovel, 2014). These findings support the idea of an oculomotor planning system highly tuned to different tasks. However, it is unknown to what degree the human brain can adapt its eye movement plans to the particular properties of the

faces presented in a task. Does the brain have access to a wide set of optimal eye movement plans to faces that it can utilize for specific facial states and tasks? Or instead, does it use a simplified set of plans matched to the more prevalent facial states encountered during everyday life? One possibility is that a simpler and less flexible heuristic strategy is used to process faces. Such a strategy might reflect the naturally occurring statistics of face stimuli that humans encounter in everyday life. In early and mid-level vision, the spatiotemporal sensitivity in the retina and early cortical areas are tuned to statistical regularities of visual information in the environment (Geisler & Ringach, 2009; Simoncelli & Olshausen, 2001). In high-level vision, the processing of the inverted faces is distinct from upright faces and related to their significantly lower statistical occurrence on human retinas through development (Belle, Graef, Verfaillie, Rossion, & Lefèvre, 2010; Farah, Tanaka, & Drain, 1995; Guo, Oruç, & Barton, 2009; Jacques, d'Arripe, & Rossion, 2007; Sekuler, Gaspar, Gold, & Bennett, 2004).

Here, we measure human eye movements during a gender identification task and investigate how the theoretical optimal and human empirical fixation strategies vary with the emotional expression of the faces. We ask whether the informative features and optimal point of fixation for gender identification vary across emotional expressions and whether the human brain takes into account this information to optimize eye movements for the task. We use the framework of a Bayesian Ideal Observer (BIO) that processes sub-regions of faces (Region of Interest Ideal Observer; ROI) to quantify the discriminatory information within each facial feature. In addition, we use a BIO that is constrained by a foveated visual input (Foveated Ideal Observer; FIO, Peterson & Eckstein, 2012) to assess how the theoretical optimal point of initial fixation for gender is influenced by the emotional expression of the

faces. Finally, we use convolutional neural network (CNN) model to simulate differences in human exposure to different statistics of facial expressions. We investigate whether these differences in exposure affect the performance of the CNN in the gender discrimination task with a different frequency of facial expressions compared to what it was exposed to.

The analysis with the ROI and FIO models shows that there is an increase in information in the mouth region in happy-expression faces for a gender identification task compared to neutral-expression faces. This leads to a shift downward in the FIO's theoretical optimal point of initial fixation as well as an increase in performance at the new optimal spot. However, our results show that humans are unable to fully take advantage of this information. We propose that observers' strategy not to use mouth expression information for gender discrimination may be related to the statistical distributions of occurrence of neutral and happy facial expressions when we first encounter an individual in an ecologically valid environment. We simulate this strategy with a CNN model and present measurements of frequencies of emotions of faces encountered in the wild captured with eyewear-embedded cameras. We propose a computational model that discounts mouth information to a large degree to account for human performance.

3.2 Materials and Methods

Human Psychophysics Studies.

Participants.

The first experimental condition, the free-viewing study, (see Procedure section below for details) was completed with a group of eighteen undergraduate students (11 female and 7 male, aged 18-22), who participated in the study for course credit. Data from three of the participants (2 female and 1 male) was not used due to them being identified as individuals who belong to a group of about 10% of the population that make the initial fixation to lower features (tip of the nose and mouth) than the other 90% of the population (M. F. Peterson & Eckstein, 2013). This left data from fifteen participants in the free-viewing condition. All of the fifteen remaining students participated in the subsequent forced-fixation experimental condition (see Procedure section below for details). Informed consent was obtained from all subjects and guidelines provided by the institutional review board of the University of California, Santa Barbara were followed.

Power Analysis.

We use a small number of participants relative to many studies in psychology. The main reason we do this is because of the time limitation of the forced-fixation condition (described below in Experimental Conditions), which requires around 300-500 trials per fixation (we use 5 fixation locations) per stimulus set (we have 2: happy and neutral faces) in order to get reliable results. However, we show using a sampling method, that the number of participants that we use is also enough to discern very small differences in preferred fixation positions, in a free-fixation condition (described below in Experimental Conditions). The reason for this is that there is a tradeoff between the variability in preferred fixation position

within a collection of trials taken from fewer participants but with lots of trials per participant, relative to more participants and fewer trials per participant.

We use a database of 285 participants who completed a free-fixation human face identification task to randomly sample with replacement the mean vertical fixation positions of 5, 15, and 30 participants and with replacement different numbers of trials per participant. For each data point we calculate the standard error of the vertical fixation position where the number of samples is the product of the number of participants and the number of trials per participant (i.e. (number of participants)x(number of trials per participant) and plot this in Figure 1 below. We repeat this procedure 100 times for each data point and plot error bars that represent the standard error of the standard errors that were found for the preferred vertical fixation position. The plot shows that a similar standard error of the mean of the preferred vertical fixation position can be achieved with fewer participants and lots of trials per participant, relative to having more participants and fewer trials per participant.

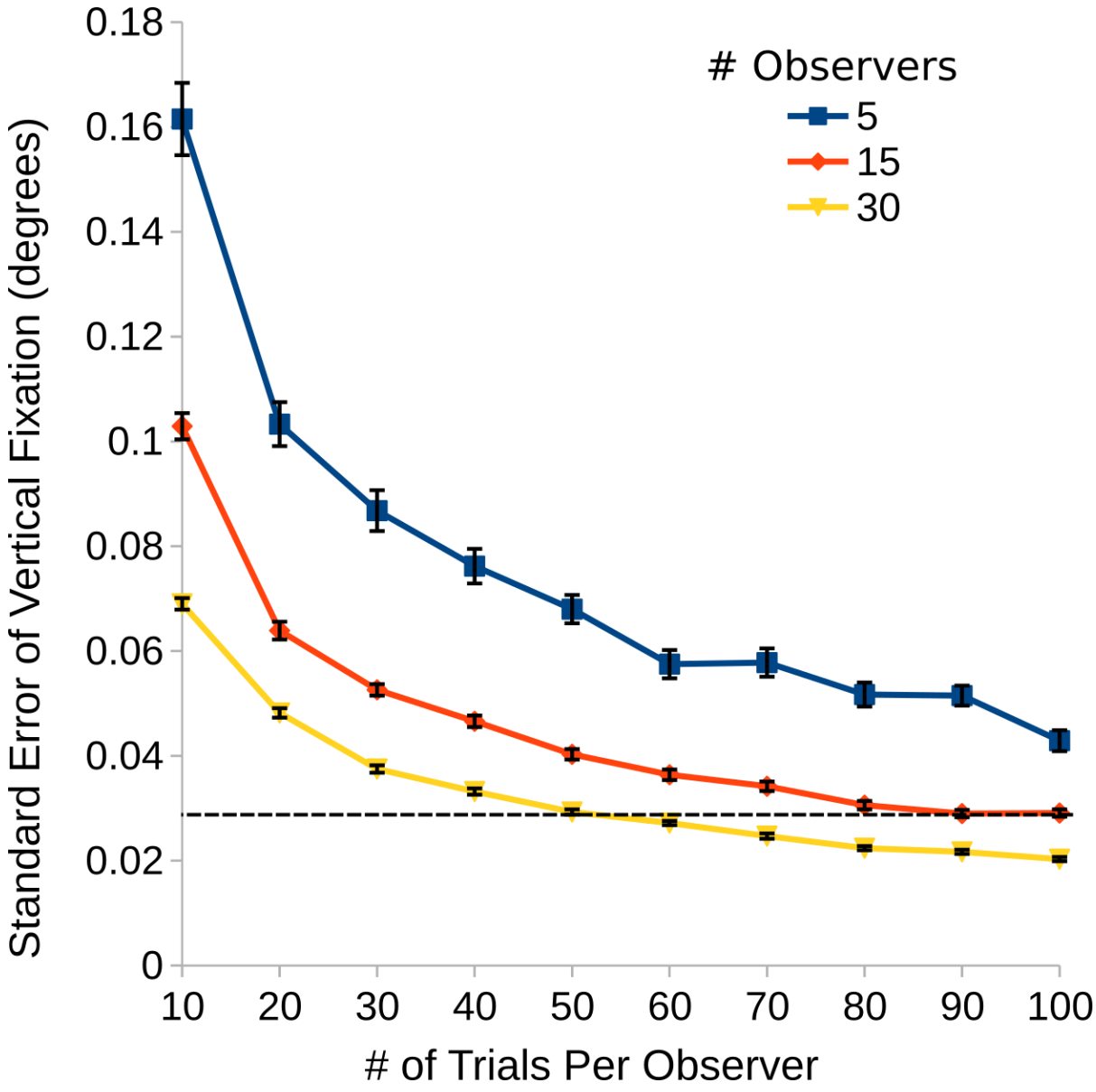

Figure 3: This plot shows the tradeoff in standard error of the mean of the preferred vertical fixation position found with different combinations of participant sample sizes and trial sample sizes per participant. We use a database of 285 participants who completed a free-fixation human face identification task to randomly sample with replacement the mean vertical fixation positions of 5, 15, and 30 participants and with replacement different numbers of trials per participant. For each data point we calculate the standard error of the vertical fixation position where the number of samples is the product of the number of participants and the number of trials per participant. The dashed line shows that a similar standard error can be achieved with 15 participants with 100 trials per participant, relative to 30 participants with 50 trials per participant.

Apparatus and Materials.

MATLAB Psychtoolbox and Eyelinktoolbox software were used to run the eyetracker from a display computer as well present visual stimuli on the display screen. The display used was a Barco MDRC 1119 monitor set to a 1280x1024 pixel resolution and was located 76.5cm away from the observer's eyes. The display was linearly calibrated with a minimum luminance of .05 cd/m² and a maximum luminance of 126 cd/m².

Eye-tracking. The left eye of each participant was tracked using an SR Research Eyelink 1000 Tower Mount eye tracker sampling at 250 Hz. A nine-point calibration and validation were run before each 125-trial session, with a mean error of no more than 0.5º of visual angle. Saccades were classified as events in which eye velocity was greater than 35º and eye acceleration exceeded 9,500º per square second. The recommended thresholds by SR for cognitive research are an eye velocity of 30º and an eye acceleration of 8,000º per square second. The minor increase of the velocity and acceleration thresholds in our parameter settings allowed us to better control the number of "broken fixations" during the initial fixation stage at the beginning of every trial prior to the presentation of the stimulus.

Stimuli.

In this experiment, observers completed a gender identification task. Eighty face images were used. The images consisted of forty identities, with two images; a neutral and happy expression for each identity. Half of the identities were male and half were female (Figure 2a). All of the images were taken with constant diffuse lighting, distance, and camera settings. A Canon digital camera was used. The digital pixel value was a non-linear saturating function of luminance (a standard Canon log-cine transfer function). The images were normalized by scaling and cropping, such that the center of the eyes was 2/5 of the

image height below the top of the image and the chin was 1/50 of the image height above the bottom of the image. The faces were luminance-mean normalized to 25 cd/m² and shown to participants at a contrast of .032 (RMS contrast of .134) and 0.044 (RMS contrast of .143) for the practice blocks), where part of that contrast variation came from added Gaussian white noise with a standard deviation of 2.75 cd/m² (corresponding to a noise RMS contrast of .11). Participants viewed the face stimuli 76.5cm away from the display resulting in a square stimulus (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. The large size of the faces, more typical of conversational distance, was chosen: (1) to allow measurements of larger variations of perceptual performance with fixation position (for small faces perceptual performance is less sensitive to fixation position within the face); (2) to allow more precise measurements of fixation positions relative to facial features. In addition, the large faces (e.g., 10 deg. width, 15 deg. height) have been shown to optimize face identification (Yang, Shafai, & Oruc, 2014).

Procedure.

Observers performed a gender identification task with 45 total blocks consisting of several different conditions. Each block consisted of 125 trials. The conditions varied depending on whether subjects were allowed to freely make saccades (free-viewing) or were forced to fixate specific locations (forced-fixation). Participants were first run through two free-viewing practice blocks with higher contrast, during which they could get familiarized with the task. Three free viewing blocks were then run with mixed happy and neutral faces (50% probability of each), neutral faces separately, and happy faces separately. The free-viewing condition was followed by a forced-fixation condition of 40 blocks with intermixed

happy and neutral faces (50% probability of each). The participants were given instructions to maximize their accuracy in the task by trying to choose the correct gender on each trial. They were also told the order of the conditions when they got to each one and received instructions on the differences between them. All 15 observers completed the free viewing blocks prior to the forced fixation conditions.

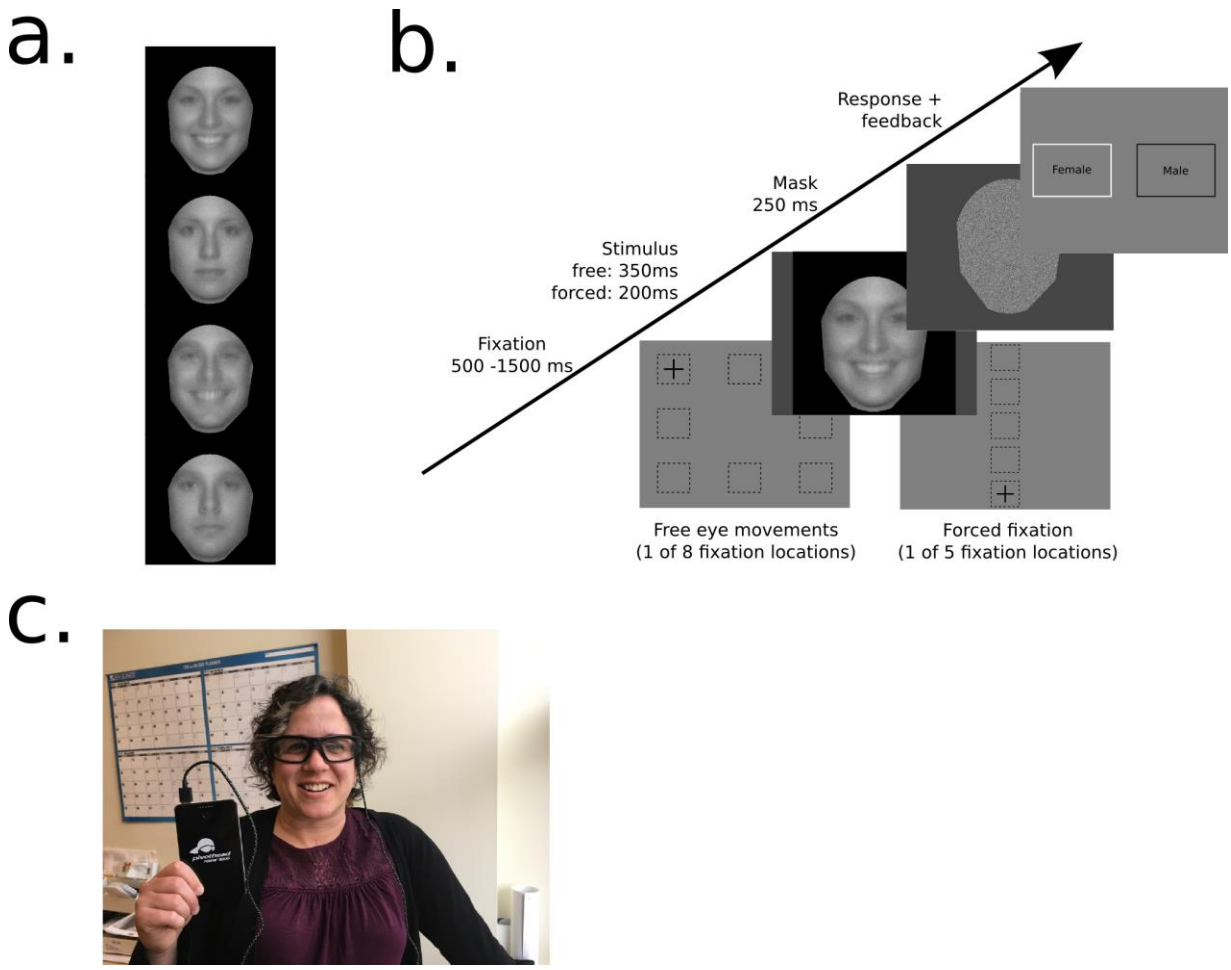

Figure 4: ***a)*** *Averages of the 4 stimuli categories are shown with happy female, neutral female, happy male, and neutral male categories from top to bottom, respectively. An average image was created by taking the mean of the grayscale luminance values across 20 individual face images in each category. (b) A trial time line is shown. In the free-viewing condition, observers made saccades to the centrally presented face from a fixation cross in one of eight randomly chosen locations. A stimulus presentation time of 350ms was used in the free-viewing condition. Separate blocks were used for happy-expression faces and neutral-expression faces in addition to one intermixed block. In the forced-fixation condition, observers fixated one of five locations (with four of the locations corresponding to the top of*

the forehead, the eyes, the nose, and the chin, and the fifth one corresponding to an individual preferred point of fixation taken from the free-eye-movement condition). A stimulus presentation time of 200ms was used for the forced-fixation condition. Here an average happy female stimulus is shown without noise and high contrast for illustration purposes (in the actual experiment contrast was substantially lower and white noise was added). At the end of each trial, participants had unlimited time to select with the mouse 1 of 2 possible genders displayed on the screen. As soon as a decision was made, feedback was given by outlining the correct gender. (c) The eyewear-embedded camera and the attached external battery used to collect data for expression frequencies in the natural world are shown.

Experimental Conditions.

During free-viewing blocks, participants started a trial by pressing the space bar while fixating a cross (.44 º x .44 º) in one of eight randomly chosen locations located on average 13.94º from the center of the stimulus. During forced-fixation blocks, the cross was located in one of 5 locations. Four of these fixation cross locations roughly corresponded to the forehead, eyes, nose, and mouth, and were placed at 5.07 º intervals along the vertical midline of the face. The fifth location was taken from each participant's individual preferred point of initial fixation in the free-viewing condition and was also restricted to the vertical midline of the face. The fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants moved their eyes more than 1º from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus. The stimulus was then shown for 350ms in the free-viewing condition, and 200ms in the forced-fixation condition. The shorter presentation time for the forced-fixation condition was used in order to account for the fact that participants did not need time to make an eye-movement from the periphery of the screen, as they did in the free-viewing condition. At the end of each trial, participants had unlimited time to select with the mouse one of two genders that were displayed on the screen without noise. As soon as a decision

was made, feedback was given by outlining the correct face. Figure 2b shows a timeline of a single trial.

Measurement of Frequency of Facial Expressions in the real world.

In order to determine adult observers' exposure frequency to various facial expressions during the course of regular daily activities we re-analyzed the footage collected as part of a previous study (Oruc, Shafai, Murthy, Lages, & Ton, 2018). This footage was collected from thirty adult participants via eyewear-embedded cameras and had been previously analyzed to examine total exposure duration to faces and frequency of various attributes such as gender, ethnicity, pose, viewing distance and familiarity. Here, we examine the exposure frequency distribution across five expression categories: neutral, happy, sad, angry, and other.

Eyewear-embedded camera.

The footage was acquired using a high-resolution 75° field-of-view eyewear-embedded camera, Pivothead Durango (http://www.pivothead.com/). The camera was set to time-lapse mode to capture still images at the rate of 1 shot / 30 s at 3-megapixel resolution. We replaced the shades with clear lenses and connected the glasses to a pocket-sized external battery (Pivothead Power Pro Refuel 8000), which the participants carried near or on their person (Figure 2c).

Participants.

Thirty adults (14 females; mean age = 31.9 ± 8.4 years, range 20-54) participated in the study. Out of the 30 participants: 28 were Caucasian, one was African and one was Asian; 18 participants recorded footage on a workday. The participants' occupations included four researchers, five engineers, one youth leader, two teachers, eleven students, one

unemployed, one accountant, one administrative coordinator, one managing director, one customer services specialist, one bookkeeper, and one lab manager. Average footage recorded per participant was just over 7 hours and 26 minutes (range: 2 h 35 m - 13 h 23 m) for a total duration of 209 hours and 57.5 minutes (25,195 frames).

Procedure.

Participants were asked to wear the recording glasses during waking hours of one day. They were instructed to turn on the camera upon waking and go about their daily activities as usual. The camera automatically captured a still image every 30 s throughout the day without any additional actions from the participants. At the end of the recording day, participants were given the option of connecting the glasses to their personal computer to review their footage in order to give them the opportunity to remove any images of a private nature (e.g. bathroom visits). Participants also completed a post-participation questionnaire where they indicated their gender, age, ethnicity, occupation, whether the recording was done on a work-day (vs. non-work day), whether they removed the glasses for any period of time, and any additional comments they had. The protocol was approved by the review boards of the University of British Columbia and Vancouver General Hospital, and informed consent was obtained in accordance with the principles in the Declaration of Helsinki.

Data coding and analysis.

The footage was pre-processed for automated detection of faces with custom Matlab scripts. A manual adjustment of automated detections was then applied, in which bounding boxes were drawn around faces that were missed by the automated process. In addition, false detections and redundant bounding boxes due to multiple automated detections of the same face were deleted such that there remained only one bounding box around every face that

was captured in the footage. Face images that appeared in media, (e.g., print, screen) were not coded. Each individual participant's footage was manually annotated for facial expression by classifying each detected face into one of five expression categories: neutral, happy, sad, angry, and other. The footage was analyzed completely by two independent coders. The coders' ability to interpret the images was improved by using the context of the situations depicted in the still images surrounding a particular detected face. Agreement between two coders was defined as the correlation coefficient between the two independent coders' frequency estimates. There was high agreement between the expression annotations of the two coders as reflected by a high correlation of .98. We averaged the ratings of the two coders to obtain a single frequency estimate for each expression category.

Ideal Observer Models.

In the following sections, we briefly describe several Ideal Observers that we use to model the face perception task presented in this paper and explain the rest of the algorithmic details in the Appendix.

Bayesian Ideal Observer.

Here, we run several different variants of an ideal observer model, starting with a standard ideal observer, which utilizes image information to achieve the highest possible performance and does not simulate the foveation of the visual system like the FIO described below. We run a face gender identification task with a set of 80 (20 male neutral, 20 male happy, 20 female neutral, and 20 female happy; the same identities are used for neutral and happy faces) front-view face images that are normalized for the position of the eyes and chin as well as for contrast (see the Stimuli subsection of Human Psychophysics Studies above for details). On each trial of the simulation, the face images $\{\mathbf{f}_1,\ldots,\mathbf{f}_{80}\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen as a signal to be presented on that trial. The same contrast and additive white noise that was used for humans is then added to a chosen template, $\mathbf{s}_i$. The input data, $\mathbf{g}$, to the ideal observer on each simulated trial is then the sum of a random (1 of 80) face template, $\mathbf{s}_i$, and external noise, $\mathbf{n}_{ex}$.

$$\mathbf{g} = \mathbf{s}_i + \mathbf{n}_{ex} \qquad (3.1.1)$$

The ideal observer does not have any sources of suboptimality such as internal noise or filtering operations on the face template, $\mathbf{s}_i$, that models foveation. Using Bayes rule, the ideal observer finds a set of posterior probabilities, one for each hypothesis that face f from

gender *g* and emotion *e* (happy or neutral) was shown, $H_{g,e,f}$, given the image data, $\mathbf{g}$. Here we use the index, *f*, to represent a calculated posterior probability for a particular face being shown, in contrast to the index, *i*, which represents the actual ground truth signal that was shown on a particular trial.

The posterior probability, $P(H_{g,e,f} \mid \mathbf{g})$, is calculated using the prior probabilities, $P(H_{g,e,f})$, and the likelihood, $P(\mathbf{g} \mid H_{g,e,f})$, of the image data, $\mathbf{g}$, given the presence of each face, f from gender *g* and emotion *e*:

$$P(H_{g,e,f} \mid \mathbf{g}) = \frac{P(\mathbf{g} \mid H_{g,e,f})P(H_{g,e,f})}{P(\mathbf{g})} \propto P(H_{g,e,f})P(\mathbf{g} \mid H_{g,e,f}) = l_f \qquad (3.1.2)$$

Then to find the posterior probability, $P(H_g \mid \mathbf{g})$, of the presence of a specific gender, the sum is found across the posterior probabilities of individual faces and across two emotions belonging to that gender:

$$P(H_g \mid \mathbf{g}) = \sum_e \sum_f P(H_{g,e,f} \mid \mathbf{g}) \qquad (3.1.3)$$

The normalizing factor, $P(\mathbf{g})$, in equation (2.1.2) is the same for all posterior probabilities, so it can be ignored without changing the result. The likelihood, $P(\mathbf{g} \mid H_{g,e,f})$, of the signal having come from a particular face is calculated from a known distribution that comes from a product of distributions of individual pixel noise (see Appendix for details).

The maximum posterior probability is then chosen as the answer at the end of a simulated trial:

$$decision = \underset{g}{\operatorname{argmax}}(P(H_g \mid \mathbf{g})) \tag{3.1.4}$$

Region of Interest Bayesian Ideal Observer.

In order to understand which regions of a face are important for this particular task we also run a Region of Interest Ideal Observer (ROI), which is a Bayesian Ideal Observer that is separately run using small sections of the face stimuli image at a time. The calculations are the same as for the ideal observer, except that in contrast to equation (3.1.1), the data, $\mathbf{g}_s$, is now the sum of a random (1 of 80) face template, $\mathbf{s}_{i,s}$, and external noise, $\mathbf{n}_{ex}$, where *i* indexes the face template and *s* indexes the section of the face for which performance is separately calculated:

$$\mathbf{g}_s = \mathbf{s}_{i,s} + \mathbf{n}_{ex} \tag{3.2.1}$$

The signal $\mathbf{s}_{i,s}$ on each simulated trial is now taken from a specific 30x30 pixel section from a randomly chosen face template, *i*. Figure 3a.1 and 3a.2 shows how small sections of a face are processed at a time and likelihoods are found for each section. Figure 3a.3 shows a performance map that is created by sampling different sections across the face stimulus. Here, we run a simulation with 30,000 trials. Due to computational constraints, we only sample the sections every 10th pixel rather than every adjacent pixel, which results in a 47x47 performance map (it is not 50x50 because of the 30px section size). This map is then resized using bilinear interpolation to a 500x500 pixel performance map to match the size of the face images.

Foveated Ideal Observer (FIO) Model.

A spatially variant contrast sensitivity function (SVCSF) was used to model the

degradation of the quality of information obtained in the periphery of a foveated visual system (M. F. Peterson & Eckstein, 2012):

$$SVCSF(f,r,\theta) = c_0 f^{a_0} \exp(-b_0 f - d_0(\theta) r^{\eta_0} f) \quad (3.3.1)$$

where *f* is spatial frequency in cycles per degree of visual angle. The terms a_0, b_0, and c_0, were chosen constants set to 1.2, 0.3, and 0.625 respectively, to set the maximum contrast at 1 and the peak at 4 cycles per degree of visual angle at fixation. The polar coordinates *r* and θ specify the distance in visual angle and direction from fixation. d_0 specifies the eccentricity factor as a function of direction, which represents how quickly information is degraded in the periphery. η_0 specifies the steep eccentricity roll off factor. In the model simulations, different parameters are used for d_0 for the vertical up, *du*, vertical down, *dd*, and horizontal, *dh*, directions. The parameters *du*, *dd* , *dh*, and η_0 are used from a previous fit with the Foveated Ideal Observer (FIO) model to match human performance (proportion correct) as a function of fixation position (5 different fixations down the vertical midline of the face) of a face identification task with a set of 20 different observers that did not participate in the experiments of this study (M. F. Peterson & Eckstein, 2012). The values used for parameters *du*, *dd*, *dh*, and η_0 respectively, are 1E-4, 2.4E-4, 5E-5, and 5. The Akaike Information Criterion (Akaike, 1974), which takes into account the variance for each data point, is used as a distance measure. The same parameters are used for the gender identification task in this experiment, except for the internal noise parameter, which shifts the performance curve downwards or upwards but does not significantly alter the shape of the curve and the relative rank order of accuracies across fixation points. The circular plots between Figure 3b.1 and

2b.2 show examples of 2d contrast

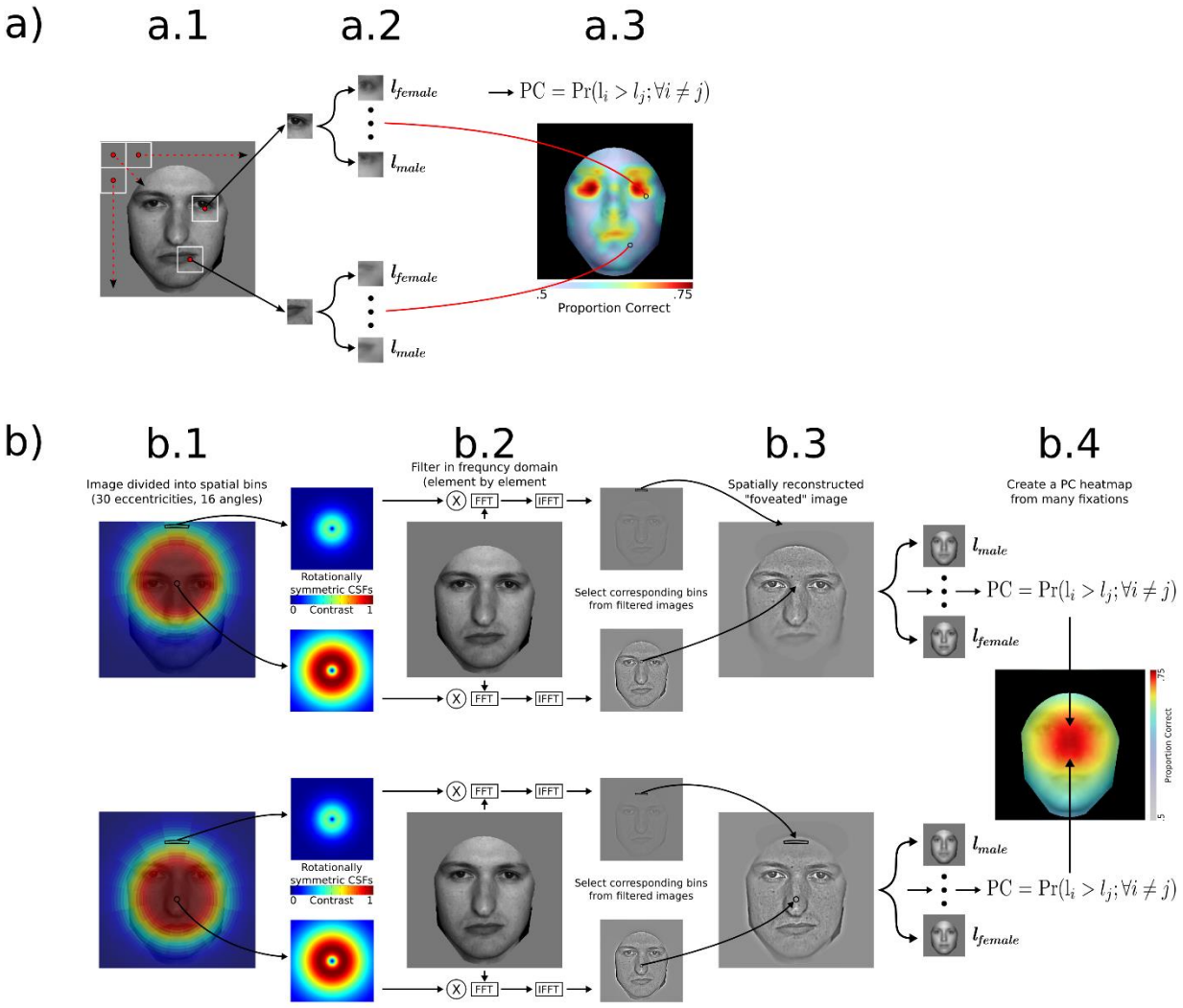

Figure 5: ***a)*** *A flow chart for a Region of Interest Ideal Observer. (a.1) An Ideal Observer is separately run for each small 30x30px section of a face image corresponding to a center point that is sampled every 10px. (a.2) On each simulated trial, likelihoods are found for a chosen face to male be or female. The likelihoods are themselves sums of likelihoods of individual faces for each gender. (a.3) The maximum likelihood principle is used to find performance in the gender task for each separate face section and output a performance map that shows which parts of a face are the most informative for this task.* ***b)*** *A summary of the process of the computations in the FIO for two fixations. The top panels show a fixation point that is below the eyes, which is optimal in several different face discrimination tasks, including gender identification with neutral faces. The bottom panels show a fixation that is at the tip of the nose, which is the theoretical optimal fixation point when viewing happy faces in a gender identification task. (***b.1-b.3)***, The filtering operation for a noiseless template.* (b.1), *A face image is conceptually divided into bins that correspond to specific CSFs as a function of retinal eccentricity. Contrast sensitivity functions that correspond to*

the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the CSF plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low-contrast blue in the CSF plots). (b.2), The image is transformed into the frequency domain, filtered separately by each possible CSF (here only two are shown), and then transformed back into the spatial domain, resulting in a set of differently filtered images corresponding to each bin. (b.3), Corresponding bins are then extracted from the filtered images and input into a composite image that simulates foveation. The procedures in b.1–b.3 are then repeated for each of the rest of the noiseless face images, as well as for the noisy input to the model on a particular trial. A set of response variables are then calculated, from which a set of likelihoods is found of each face given the noisy image input. **(b.4)***, A decision of which face was shown is made by taking the maximum likelihood. Across many trials, a set of proportion correct (PC) values is found, one for each fixation point, and then combined into a heatmap. iFFT, Inverse FFT.*

sensitivity functions at 2 different locations with respect to the fixation position. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low contrast in blue in the plots).

Here, we simulate the same face gender identification task with a set of 80 face images. On each trial of the simulation, the face images $\{\mathbf{f}_1,\ldots,\mathbf{f}_{80}\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen. The same contrast and additive white noise that was used for humans is then added to a chosen template, *i*. However, in contrast to equation (2.1.1), the noisy template is then linearly filtered with the SVCSF and corrupted with additional internal white noise to become the input data, $\mathbf{g}_k$, to the foveated ideal observer:

$$\mathbf{g}_k = \mathbf{E}_k(\mathbf{s}_i + \mathbf{n}_{ex}) + \mathbf{n}_{in} \tag{3.3.2}$$

where $\mathbf{n}_{ex}$ is the external Gaussian white noise, $\mathbf{n}_{in}$ is the internal Gaussian white

noise, and $\mathbf{E}_k$ is the linear operator that simulates the fixation dependent foveation of the input. This foveated signal is compared (by taking a dot product) to similarly foveated noiseless templates (original face images) to arrive at a set of responses, $\mathbf{r}_{f,k}$, which come from a multivariate Gaussian distribution with a known mean, $\boldsymbol{\mu}_{f,k}$, and covariance matrix, $\sum_k$ (see Appendix for details on how they are calculated):

$$\mathbf{r}_{f,k} \sim MVN(\boldsymbol{\mu}_{f,k}, \sum_k) \tag{3.3.3}$$

Using Bayes rule, the FIO finds a set of posterior probabilities, one for each hypothesis that face *f* from gender *g* and emotion *e* (happy or neutral) was shown, $H_{g,e,f}$, given a set of responses $\mathbf{r}_{f,k}$. The posterior probability, $P(H_{g,e,f} \mid \mathbf{r}_{f,k})$, is calculated using the prior probabilities, $P(H_{g,e,f})$, and the likelihood, $P(\mathbf{r}_{f,k} \mid H_{g,e,f})$, of the set of responses given the presence of each face, *f*, and the observer's fixation at spatial location, *k*:

$$P(H_{g,e,f} \mid \mathbf{r}_{f,k}) = \frac{P(\mathbf{r}_{f,k} \mid H_{g,e,f})P(H_{g,e,f})}{P(\mathbf{r}_{f,k})} \propto P(H_{g,e,f})P(\mathbf{r}_{f,k} \mid H_{g,e,f}) \tag{3.3.4}$$

The normalizing factor, $P(\mathbf{r}_{f,k})$, in equation (2.3.4) is the same for all posterior probabilities, so it can be ignored without changing the result. Then to find the posterior probability, $P(H_g \mid \mathbf{r}_{f,k})$, of the presence of a specific gender, the sum is found across the posterior probabilities of individual faces and across two emotions belonging to that gender.

$$P(H_g \mid \mathbf{r}_{f,k}) = \sum_e \sum_f P(H_{g,e,f} \mid \mathbf{r}_{f,k}) \tag{3.3.5}$$

The maximum posterior probability is then chosen:

$$decision = \underset{g}{\operatorname{argmax}}(P(H_g \mid \mathbf{r}_{f,k})) \tag{3.3.6}$$

Convolutional Neural Network Model.

Stimuli.

Training.

The training stimulus set was taken, with permission, from a combination of the Multimedia Understanding Group (MUG) faces database (Aifanti, Papachristou, & Delopoulos, 2010), the Cohn-Kanade (CKPlus) faces database (Kanade & and, 2000; Lucey et al., 2010), a Binghamton University faces database (L. Yin, Chen, Sun, Worm, & Reale, 2008), and an in-house faces database. The first three databases mentioned above contain movies of facial expressions starting from a neutral expression and unfolding into one of several different emotions, all of which were discarded, except for happiness. The remaining happy-expression movies were then used to manually extract a single frame of a neutral expression from the beginning of each movie and then a single frame of a happy expression from a later time point. This allowed for the creation of two datasets, one with a neutral-expression, and another with a happy expression, with the same identities in each dataset. The datasets were then further manually trimmed to discard male identities that had excessive facial hair, which may have biased the CNN model to use that as a major feature in the gender discrimination task. This resulted in a set of 137 female faces and 84 male faces for each facial expression (neutral and happy), with the same identities. In order to avoid biasing the model toward female faces, a random subset of the male faces were oversampled without replacement to create an equal set of 137 faces for each gender (274 total for each facial expression). Due to the small number of faces in this set, a data augmentation step was then implemented, where 20 copies of each face in the dataset were created, with random Gaussian white noise added to each copy. This resulted in a set of 5080 faces for each facial

expression (2540 for each gender within a facial expression dataset), where none of the faces had the exact same pixel values due to the added noise. We trained the model separately in three different ways: 1) using a neutral-expression dataset; 2) using a happy-expression dataset; 3) using a combination of mostly (87%) neutral-expression faces and a small proportion (13%) of happy-expression faces. The expression proportions for the third training set were obtained from frequencies of facial expressions found in the real world (see *Measurement of Frequency of Facial Expressions in the real world* section of the Methods above).

Testing.

The base dataset used for the test stimuli were the same 80 faces (20 male happy, 20 female happy, 20 male neutral, 20 female neutral) that were used to the run the main analyses with the ROI and FIO models as well as the human psychophysics experiments. This resulted in a dataset with an equal number of faces for each gender and facial expression, and the same identities across the two expressions. In order to get a large enough dataset to measure more precise performance differences between differently trained CNN models, a data augmentation step was then implemented, where 10 copies of each face in the dataset were created, with random Gaussian white noise added to each copy. This resulted in a set of 800 faces (2 datasets with 400 faces for each facial expression). These two datasets were then separately used to test three differently trained models (described in the *Training* section above), resulting in 6 combinations of testing and training that we describe in the results section.

Architecture and Settings.

Original resnet-18.

We use an 18-layer resnet-18 (He, Zhang, Ren, & Sun, 2015) architecture (Figure 4a) to run a 2-class gender discrimination task with the datasets described above. The network is made up of 4 "residual blocks," each of which contain 2 pairs (this number is higher for other variants of this network structure) of the same layer structure (same size and depth of feature maps). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem.

We use mini-batch (200 images per batch) stochastic gradient descent (SGD) along with a cross-entropy loss function to optimize the parameters in the model. We use hyperparameter settings of 5e-4 for the learning rate and .9 for momentum. Although this network can theoretically be run with any input image size, here we run it with an image size of 112x112 pixels due to a limitation of the resolution of the training images we used. Upscaling the training images would only increase computation time without improving performance.

Modified resnet-18 and Class-Specific Activations Visualization.

In addition to the original resnet-18 network, we use also the methodology of (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) and run a modified version of the same network in order to be able to construct a visualization of the important features in the input stimuli that are used by the network to do the gender classification task. This is done by mapping a weighted linear combination of the 14x14 feature maps of the last convolutional layer of the network onto the original 112x112 input images. The weights used to combine the feature maps come from the learned connections between the Global Average Pooling (GAP) layer,

which acts as a unidimensional representation of the 14x14 feature maps preceding it, and the class scores output by the network. For each of the 2 classes, a specific Class Activation Map (CAM) is found by using the weights connecting the GAP layer to a specific class. Although (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) used this method for localization of objects in complex classification tasks with a large number of classes, it is still useful for our purpose of visualizing the features of faces that are most discriminative for the network during this task. Since the faces are aligned during both the training and testing phases, the discriminative features should be located in specific areas across CAMs. We average the visualizations across CAMs to get a single visualization map for each testing set to get an overall representations of which face features the network is able to use the most during this task.

Figure 4b shows the modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the third and fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering the stride from 2 to 1 during the convolution operation before the last 2 residual blocks. The difference in the modified network is outlined in red.

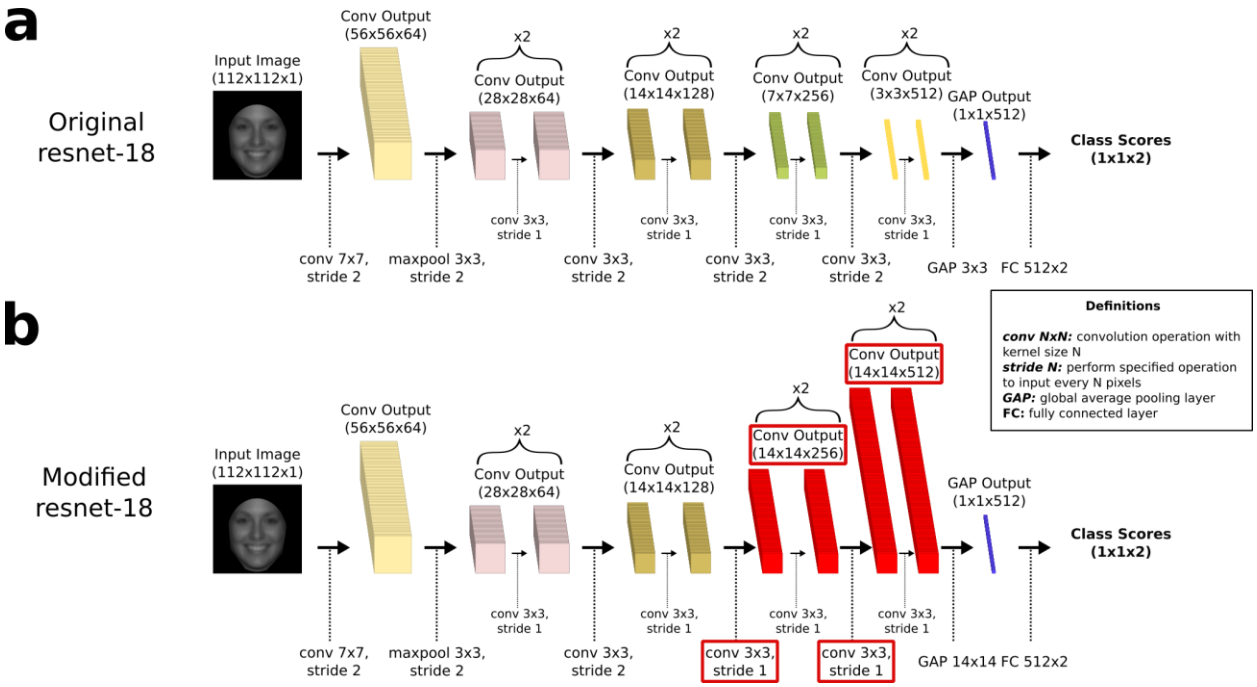

*Figure 4: A flowchart of the operations and outputs involved in the CNN network that we use. (**a**) The top flowchart shows the structure and operations involved in the original resnet-18 network* (He, Zhang, Ren, & Sun, 2015). *Although this network can theoretically be run with any input image size, here we run it with an image size of 112x112 pixels and show the sizes of feature map outputs after max pooling and convolution operations along with the chosen depths of the feature maps at each layer, which are fixed parameter settings. Similarly, although the network is able to learn to classify an arbitrary number of classes, here we show an output of class scores for a 2-class gender discrimination ask. One aspect of the resnet network that isn't explicitly shown in the flowchart is the "skip-connections" between layers of the same size. The network is made up of 4 "residual blocks," each of which contain 2 pairs of the same layer structure (same size and depth of feature maps). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem. (**b**) The bottom flowchart shows a modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the third and fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering the stride from 2 to 1 during the convolution operation before the last 2 residual blocks. The difference in the modified network is outlined in red. We implement this modification in order to output a set of 14x14 pixel feature maps instead of 4x4 pixel feature maps. This allows us to use the methodology of* (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) *to construct a visualization of the important features in the input stimuli that are used by the network to do the gender classification task. This is done by mapping a linear combination of the 14x14 feature maps onto the original 112x112 input images.*

3.3 Results

Region of Interest Ideal Observer and CNN Model show an increase of information in the mouth region of happy vs. neutral expression faces for a gender identification task.

In order to understand if the distribution of discriminative information in the face stimuli for a gender identification task changes when using happy-expression vs. neutral-expression faces, we first ran a Region of Interest Ideal Observer (ROI) analysis, which shows the most informative regions of a face for a particular task (see Methods and Figure 3a for details about the ROI). Figure 5a shows the results of the ROI separately for trials where neutral face stimuli were shown (left panel) vs trials where happy face stimuli were shown (right panel). Although the eyes are important features for both sets of face stimuli, there is an increase of information in the mouth region for happy faces, suggesting that there are gender differences in happy expressions that might arise from gender differences in expression of happiness. There is evidence that there are differences in muscle activation around the mouth region between men and women during positive expressions (smiling), with women having a greater activation and a more exaggerated smile (Soussignan et al., 2013). Figure 5b shows the corresponding results of the performance of an Ideal Observer, which uses all of the information in the faces to do the gender discrimination ask. The Ideal Observer performs better with happy-expression faces relative to neutral-expression faces because of the extra gender information in the mouth region of happy-expression faces.

In addition to running ideal observer models, we also run a CNN model that is able to simulate the development of an internal face representation based on exposure to a specific

training set, which an Ideal Observer is not able to do. Unlike an Ideal Observer, which has a perfect pixel-level representation of each faces used in this task, a CNN model is able to learn a complex feature representation of the faces. We train the model on a gender discrimination task with a stimulus set of neutral-expression faces and then separately train another model on a stimulus set with happy-expression faces. Both sets have the same face identities. Then we test the models with a separate set of neutral-expression faces and a separate set of happy-expression faces, respectively, which are the same stimuli that were used in human psychophysics experiments above. In this way, we train and test a model on neutral-expression faces, and then separately train and test a model on happy-expression faces and compare the results to the output of the Ideal Observer and ROI. Figure 5c shows visualizations of which parts of a face a CNN model uses the most (in terms of magnitude of neuron-like node activations in the last convolutional layer) to do the gender discrimination task with a model trained and tested on neutral-expression faces, and a separate model trained and tested on happy-expression faces. The model trained and tested on happy-expression faces has high activations across a large area of the face, including the eye and mouth region. In contrast, the model trained and tested on neutral-expression faces has high activations across a smaller area of the face that does not include the mouth region. These results are similar to the results of the ROI maps, which show that there is an increase in gender information in the mouth region of happy-expression faces. Figure 5d shows the corresponding results of the performance of each of the two CNN models. The model trained and tested with happy-expression faces performs better relative to one trained and tested with neutral-expression faces because of the extra gender information in the mouth region of happy-expression faces. Based on the visualizations, the model is able learn to use this

information when trained with happy-expression faces, and then is able to extract it when tested on happy-expression faces. However, model that is trained on only neutral-expression faces is unable to learn to extract this information.

Foveated Ideal Observer predicts different optimal point of fixation for happy vs. neutral expression faces for a gender identification task.

We evaluated the theoretical predictions for an optimal point of fixation for an FIO model, which took into account the foveated nature of the visual system. The model incorporated a spatially variant filtering of the visual input and integrated the information in face images optimally to compute posterior probabilities and make trial to trial decisions about the gender of the face presented (see Figure 3b and Methods for mathematical details). The parameters for the eccentricity dependent contrast sensitivity function of the FIO were obtained by fitting the FIO model to an independent data set of 20 different observers participating in a forced-fixation experiment for a face identification task (M. F. Peterson & Eckstein, 2012). Given that there is more gender information in the mouth region for happy faces, as shown with an ROI and CNN model (Figure 5a-d), the FIO (Figure 5e) predicts a downward shift in the theoretical optimal point of fixation for trials with happy-expression faces (right panel) vs trials with neutral-expression faces (left panel).

The results show that using faces with a happy expression in a gender identification task, alters the original theoretical optimal point of fixation (just below the eyes) to one that is at the tip of the nose. In order to make sure that these differences can be generalized and are not simply a reflection of the specific set of 80 faces that we used for this task, we ran the FIO analysis on another set of 80 faces and found the same results (proportion correct values from the FIO were used at 50 equally spaced points between the forehead and the mouth on

the vertical midline of the face to find an RMS of .012 for a happy-expression faces comparison, with the largest difference between two points being .018, and an RMS of .013 for a neutral-expression faces comparison, with the largest difference between two points being .02).

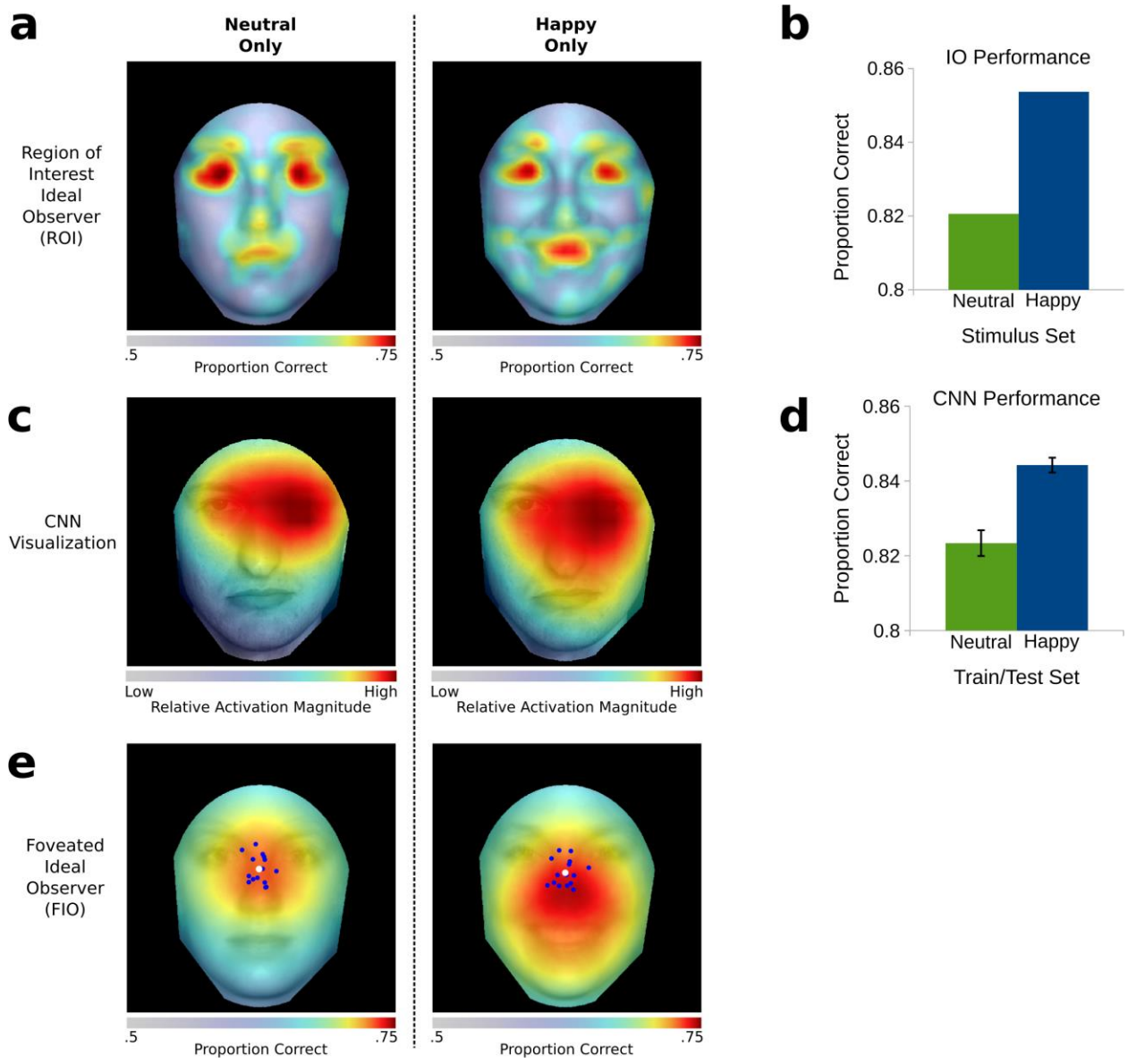

Figure 5: ***a)*** *The top row shows the results of a Region of Interest Ideal Observer performance (proportion correct) map in a gender identification task for many simulated trials where the stimulus was a neutral face (left), or a happy face (right). The ROI map for happy faces shows that there is more discriminative information in the mouth region compared to the ROI map for neutral faces.* ***b)*** *The corresponding results of the performance*

of an Ideal Observer, which uses all of the information in the faces to do the gender discrimination ask, are shown. The Ideal Observer performs better with happy-expression faces relative to neutral-expression faces because of the extra gender information in the mouth region of happy-expression faces. ***c)*** *The middle row shows visualizations of which parts of a face a CNN model uses the most (in terms of magnitude of neuron-like node activations in the last convolutional layer) to do the gender discrimination task with a model trained and tested on neutral-expression faces, and a separate model trained and tested on happy-expression faces. The model trained and tested on happy-expression faces has high activations across a large area of the face, including the eye and mouth region. In contrast, the model trained and tested on neutral-expression faces has high activations across a smaller area of the face that does not include the mouth region.* ***d)*** *The corresponding results of the performance of each of the two CNN models are shown. The model trained and tested with happy-expression faces performs better relative to one trained and tested with neutral-expression faces because of the extra gender information in the mouth region of happy-expression faces. Based on the visualizations, the model is able learn to use this information with happy-expression faces.* ***e)*** *The bottom row shows a performance map of an FIO on trials where the stimulus was a neutral face (left), or a happy face (right). The FIO for happy faces shows that there is a shift downward in the theoretical optimal point of fixation from one that is below the eyes for neutral faces to new point that is at that the tip of the nose. In addition, there is an increase in performance at the theoretical optimal point for happy faces vs neutral faces. Individual observers' average (across trials and blocks) initial fixation positions from the free-viewing condition are overlaid in blue and the white point is the average across observers (average of blue points). Performance between the neutral and happy maps of the ROI can be directly compared, as well as performance between the neutral and happy maps of the FIO. However, performance cannot be directly compared between the FIO and ROI. A noise parameter was fit to each model separately because of large differences in efficiency between the two models compared to an ideal observer. Using the same value for both would result in ceiling or floor effects in performance.*

Human Perceptual Performance and Eye movements with Happy and

Neutral Expression Faces.

Based on the theoretical modeling results, we hypothesize that if humans integrated all the information across the face optimally and adapted an optimal eye movement strategy to the specific emotional expression of the stimuli, they should also show a downward shift in their preferred point of fixation for faces with happy expressions.

A group of 15 observers participated in a face gender identification task with two experiments. The first experiment was a free-viewing condition that was used to evaluate whether the facial expressions altered the human initial preferred point of fixation. The second experiment was a forced-fixation condition that was used to evaluate identification performance as a function of fixation position. The free-viewing experiment was run in the first 3 blocks (125 trials per block), followed by 40 blocks (125 trials per block) of the forced-fixation experiment. Many more blocks were used for the forced-fixation position compared to the free-viewing study due to the need to collect performance data for 5 fixation points. See Figure 2 for a task timeline.

Free-Viewing. In the free viewing study, participants were able to make free eye movements to faces in three different experimental blocks that differed in the presentation of the happy or neutral facial expressions. In the first block, both happy and neutral-expression faces were intermixed across trials and randomly sampled with equal probability. In the second and third blocks, neutral-expression and happy-expression faces were presented separately, respectively, for 125 trials each. We used a short presentation time of 350ms to assess a preferred fixation location for a single initial saccade. The was no statistically significant difference between the preferred 1st vertical fixation position in the block where happy and neutral faces were intermixed across trials vs. the block with only neutral-

expression faces, t(14) = 1.146, p = 0.267, one-tailed. When the facial expressions were blocked, there was a small but statistically significant difference in initial vertical fixation position between the block with neutral-expression stimuli vs. the block with happy-expression stimuli, where the fixation for happy-expression stimuli was .254 degrees lower, t(14) = 3.016, p = 7.787E-3, one-tailed, corresponding to 17% of the distance between the theoretical optimal point for neutral-expression faces and the theoretical optimal point for happy-expression faces, as well as 8.6% of the distance between the eyes and nose.

Figure 5e shows observers' average initial fixation positions overlaid onto predictions of the FIO. On their first saccade, fifteen of the original eighteen observers tended to fixate a region slightly below the eyes which is consistent with the previous results from (M. F. Peterson & Eckstein, 2012), showing that this behavior is observed in about 85-90% of subjects. There was also no observed significant difference between the mean fixation position in the neutral-expression block and the theoretical optimal point of fixation for the FIO with neutral face stimuli, t(14) = 0.105, p = .918, two-tailed, corresponding to 0.026 degrees. However, there was a significant difference between the mean fixation position in the happy-expression block and the theoretical optimal point of fixation for the FIO with happy face stimuli, t(14) = 5.376, p = 9.775E-5, two-tailed, corresponding to 1.252 degrees and 85% of the distance between the theoretical optimal point for neutral-expression faces and the theoretical optimal point for happy-expression faces, as well as 42% of the distance between the eyes and the nose.

Forced-Fixation Study. The free-viewing study showed a small effect of facial expression on the initial fixation position (0.254 degrees difference). However, the FIO showed a significant difference (1.48 degrees) between the theoretical optimal point with

neutral-expression faces (just below the eyes) and the happy-expression faces (nose tip). This discrepancy with the FIO model might suggest an inability of humans to learn the new optimal point of fixation. However, it may also be the case that the FIO does not accurately predict the human optimal point of fixation for happy faces. To test this possibility, we assessed whether there was a difference in identification accuracy with neutral and happy face stimuli at 5 different points of fixation down the vertical midline of the face (four of them roughly corresponding to the forehead, eyes, tip of the nose, and mouth; the fifth was an individual preferred point of fixation found during the free-viewing condition) by forcing observers to fixate each position during the duration of the trial using only the 200ms presentation time (with feedback). We found that the FIO model fit the human forced-fixation data well for neutral stimuli. The theoretical optimal point of fixation predicted by the FIO matched the location of the empirically found optimal point in humans, which is their preferred point of fixation as seen in Figure 6a. The increase in performance at the preferred point was higher (by 1.2%) but not statistically different from performance at the eyes ($t(14) = 1.694$, $p = 0.112$, one-tailed). However, performance at the preferred point was significantly higher than at the nose ($t(14) = 3.45$, $p = 3.9E\text{-}3$, one-tailed). In contrast to the ability of the FIO to predict human performance as a function of fixation point for the neutral-expression faces, we found that there was a disagreement between human forced-fixation data and the FIO using happy-expression faces. Human performance with happy-expression stimuli is relatively flat between the eyes, preferred point, and the nose ($t(14) = 1.915$, $p = .076$, one-tailed, eyes vs. preferred point; $t(14) = 1.105$, $p = 0.288$, one-tailed, preferred point vs. nose), rather than having a shift downward of the optimal fixation point to the nose and a significant increase in performance at that point (4% difference), as predicted

by the FIO (Figure 6a).

To quantitatively compare how human and FIO performance varied with fixation position, we varied the internal noise in the FIO model to degrade its performance in order to fit the human data (minimize the Akaike information criterion (AIC) (Akaike, 1974)). All parameters related to the contrast sensitivity function remained the same as in (M. F. Peterson & Eckstein, 2012), based on fitting an independent set of observers participating in a forced-fixation face identification study (see Methods). The continuous line in Figure 6a shows the predicted model performance of FIO for happy and neutral face stimuli separately. The FIO does not successfully predict human performance with happy-expression stimuli. Note that the level of internal noise shifts the accuracy curve downwards or upwards but does not significantly alter the shape of the curve and the relative rank order of accuracies across fixation points.

An additional discrepancy between the FIO and the measured human performance is that the model predicts higher performance with the happy-expression faces (see peak performances in Figure 6a) relative to neutral-expression faces, while human gender discrimination performance is similar for both sets of emotion expressions. We also measured the efficiency (see Methods for details of this metric) of human observers relative to both an Ideal Observer model and an FIO model for happy-expression and neutral-expression trials separately. We used an average of human performance from their preferred points of fixation in the forced-fixation condition. For the FIO, the fixation position used to calculate performance was taken from the optimal point for happy-expression stimuli and neutral-expression stimuli separately. In addition, the internal noise parameter in the FIO was set to zero to have a more accurate comparison with the ideal observer. The efficiency of

humans relative to the ideal observer for neutral-expression faces and happy-expression faces is .0027 and .0022, respectively, which shows that humans are less efficient at using the extra gender information contained in happy-expression faces. Similarly, the efficiency of humans relative to the FIO for neutral-expression and happy-expression faces is .01 and .007, respectively, with humans again being less efficient in using the information in happy-expression faces. For both sets of stimuli, human efficiency relative to the FIO is higher than the efficiency relative to the ideal observer because the FIO has a major source of suboptimality that limits its own use of the information contained in the original face images. These results suggest that the human strategy might depart in a fundamental way from the FIO for gender discrimination with happy-expression faces. The small differences between the eye movement strategies and perceptual performance in discriminating gender for happy-expression faces vs. neutral-expression faces suggests that human observers might adapt a single strategy for both stimulus types. One possibility is that human strategy is optimized for the most frequent type of facial expression that occurs in the natural world when we first encounter or come across a person and make an initial eye movement to that person's face. In the next section, we investigate this hypothesis with experimental measurements and computational modeling efforts.

Human eye movement strategy is adapted to the statistical occurrence of emotional expressions in the real world.

To investigate the possibility that human eye movements and perceptual strategy might be related to the naturally occurring statistics of emotion expressions in the real world, we first analyzed the real-world facial expression frequency data collected with a mobile eyewear-embedded camera (see "Measurement of Frequency of Facial Expressions in the real world" part of Methods section). We then evaluated several computational models that implemented components that are conceptually consistent with the findings of the facial expression frequency data.

Facial Expression Frequency Analysis Shows a Low Frequency of Happy vs. Neutral Expressions.

A group of 30 participants wore a mobile eyewear-embedded camera as they went about their daily lives and came into contact with other people throughout the day. Out of a total of 25,195 frames collected, 7641 faces were detected in 4940 frames. An average of 254.7 (SD=148.8) faces per participant were detected. As shown in Figure 6b, the results indicate that the overwhelming majority of face exposure is to neutral faces. A repeated-measures ANOVA revealed a significant main effect of expression category (F (4, 112) = 1804.27, $p << .001$). Post hoc pairwise comparisons showed that frequency of neutral expression (M = 85.52%, SD = 9.19) was significantly greater than that of happy (M = 10.25% SD = 6.62), sad (M = .25% SD = .45), angry (M = .33% SD = .69) and other (M = 3.67% SD = 3.68) expressions based on Tukey-Kramer Multiple-Comparison Tests (all p's < .05). In addition, frequency of happy expression was significantly greater than those of sad, angry and other expression categories (all p's < .05). Finally, the other category frequency

exceeded that of sad but not angry expressions. No differences were found between sad and angry expression categories ($p > .05$). Overall, the most frequently seen basic expression is happy, though, it accounts for only a small fraction (approx..10%) of total exposure. Sad and angry expressions are rarely encountered. The "other" category was an umbrella term that included all subtle expressions and un-coded basic expressions, such as surprise, yet accounted for only 4% of total exposure. Thus, our analysis revealed that in the naturalistic setting exposure to overtly expressive faces is infrequent. Rather, faces encountered on a daily basis predominantly display a neutral expression.

Modification to the FIO that takes into account differences in neutral and happy expression frequencies.

Here we implement various modifications of the FIO model trying to incorporate components related conceptually to the naturally occurring statistics of emotional expressions. We evaluate two different versions of the FIO model. In the first version, we use a model that applies prior probabilities to the representations of happy-expression faces and neutral-expression faces based on the measured statistics of the expressions. In the second version, we assume that the human perceptual strategy is adapted to neutral-expression faces and relies less on the mouth (relative to the optimal observer) during a gender discrimination task. Below, we describe and assess both models.

FIO with prior probabilities of expressions.

In the FIO model, the prior probabilities represent the proportion of time that either a happy or neutral facial expression appears on a particular trial. These probabilities are part of the calculation that the model uses to find which category a stimulus likely belonged to. Running a standard FIO with uniform priors (equal probabilities) for happy-expression and

neutral-expression faces leads to significant differences in performance as well as a change in the theoretical optimal point of fixation between them, neither of which we observe in humans (Figure 6a). We first attempt to explain this discrepancy by running the same FIO model, except with a significantly lower prior probability for happy expressions and a higher prior probability for neutral expressions. Although we change the priors in the calculation of the posterior probabilities, we still keep the actual frequency of happy and neutral faces presented to the model at 50%, and then separate out trials for each of the two expression categories. We use the data on facial expression frequencies, collected in naturalistic settings, to set priors for happy and neutral facial expressions. Due to the very low frequencies of facial expressions that are not happy or neutral, we ignore them and normalize the sum of happy and neutral frequencies to a probability of one. This gives us priors of .893 and .107 for neutral and happy facial expressions, respectively. In this way, even though the actual probabilities of happy and neutral facial expressions that are shown to the model are the same, we simulate a change in the internal expectations of the probabilities in humans by changing the priors in the model and basing them on the empirically observed frequencies of the expression categories.

Figure 6c shows the results of this model with the skewed priors (prior of .107 for happy expressions, and .893 for neutral expressions) compared to humans. Although this model results in an overall decrease in performance for happy-expression faces, this performance is now much lower than for neutral-expression faces, rather than being very close, as it is in humans. This

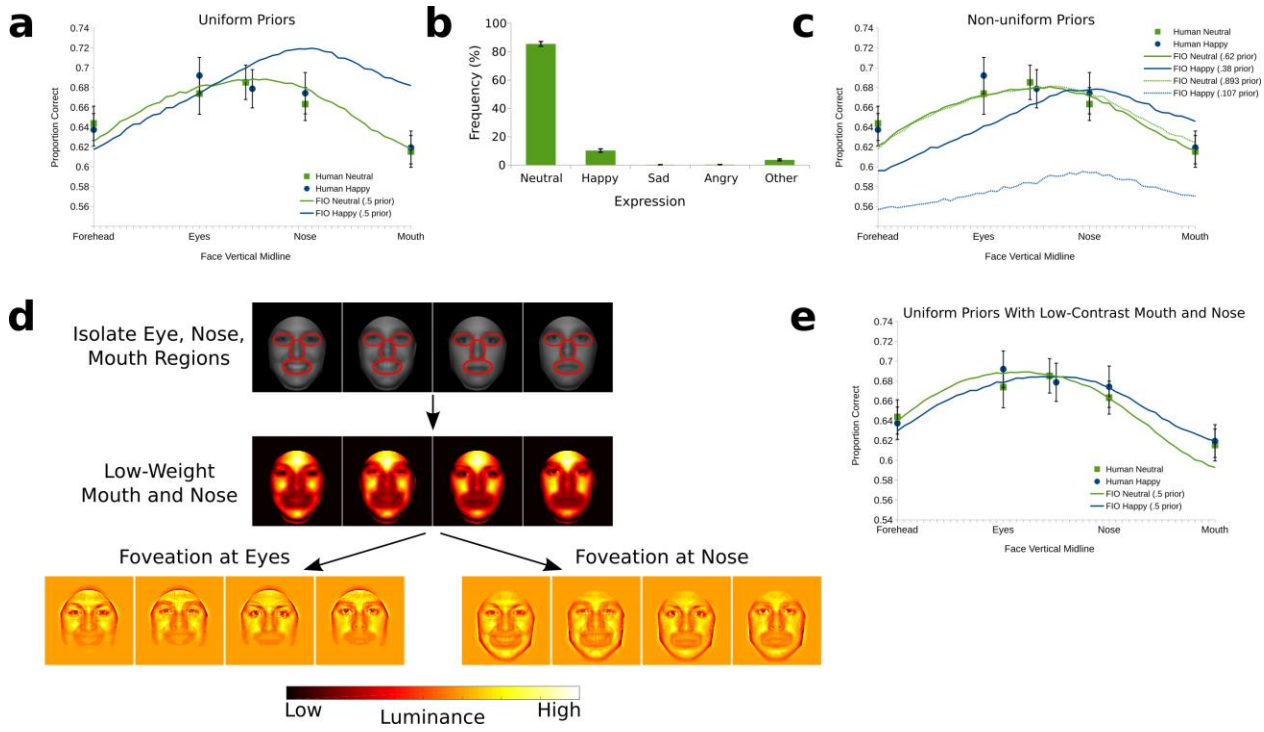

*Figure 6:**a)** Performance in the gender identification task at locations down the vertical midline of the face is shown for the FIO model with uniform priors for happy face stimuli (blue) and neutral face stimuli (green). Corresponding human performance from the forced-fixation condition is shown for happy face stimuli (blue circles) and neutral face stimuli (green squares). The middle point from the human data is misaligned between neutral and happy face stimuli because it is taken from a preferred fixation position in the free eye movement condition for each stimulus set individually. The noise in the FIO model is adjusted to best fit the model and human performance for the neutral stimuli. The FIO with happy stimuli using the same noise, shows an increase in performance and a shift downward in the theoretical optimal fixation position. **b)** We show the results of the frequency of facial expressions that were measured with eyewear-embedded cameras by multiple participants in the real world. **c)** Here, we adjust the priors for the FIO model in order to try to account for the differences in performance and the optimal point between the human data and the FIO with the happy stimuli. The noise in the FIO model is again adjusted to best fit the model and human performance for the neutral stimuli for each of the priors. When a very low prior for happy stimuli is used, performance drops substantially, but the theoretical optimal point remains unchanged. **d)** In order to account for the lack of movement in the theoretical optimal point of fixation for the happy stimuli in humans, we use a model that has a lower contrast representation of the mouth and nose region for all of the face. The damping parameters used for the mouth and nose were found with a search across 125 combinations of three contrast parameters for the eye region, nose region, and mouth region. The top panel shows the original stimuli and the middle panel shows how they are altered to lower the contrast of the mouth and nose region in this model. The bottom panel shows what the stimuli look like after foveation at two example positions, the eyes and nose. These visualizations show how the FIO model internally represents the dampened face stimuli at different fixation positions. For a foveation position at the eyes, the lower part of the face*

that includes the nose and mouth is already processed with low resolution, so the dampened nose and mouth regions are not expected to affect performance with happy-expression faces. However, for a foveation position at the nose, where the nose and mouth are processed with high resolution, those regions are now less informative because they are dampened. ***e)*** *Here the performance is shown for humans and the FIO model with uniform priors that uses altered internal face representations with a lower contrast mouth and nose region. The performance profile of the FIO with altered internal face representations during trials with happy stimuli is much closer to the human performance profile with happy stimuli compared to the original FIO model.*

model was fit to human neutral-expression data, with an AIC measure of .551. In comparison, the AIC for happy-expression human data compared to the corresponding results with the model is much higher at 22.962. A possibility is that the ideal observer's perfect representations of the set of faces as well as optimal integration of the facial features makes the FIO's performance very sensitive to the priors. In comparison, humans have intrinsic uncertainty in the representations of the faces, as well as suboptimal integration of the features (Gold et al., 2012). In order to compensate for these effects, next we evaluated a model for which the priors were fit so that it would have the same peak performance for the happy-expression and neutral-expression face stimuli. Figure 6c shows the results of this model with the resulting priors (prior of .38 for happy expressions, and .62 for neutral expressions) compared to humans, along with the results of the original skewed priors (prior of .107 for happy expressions, and .893 for neutral expressions) for comparison.

Note that the noise parameter for both of the models was fit to match human performance for the neutral-expression faces. Although the new model with the priors as free parameters results in the same peak performance for the happy-expression and neutral-expression faces, the optimal point of fixation predicted by the model remains unchanged (relative to the more extreme priors) and a large discrepancy remains with the human accuracy as a function of point of fixation (forced-fixation performance). Although this

model is a better fit to the happy-expression data (AIC of 3.53 compared to 22.962 for the more skewed priors), the fit is still significantly worse than to the neutral-expression data (AIC of .381).

Gender face representation that down-weights the lower part of the face.

The second model we investigate assumes that for a gender discrimination task, humans adopt a feature strategy that is optimal for the most frequent emotion expressional (neutral) by down-weighting the use of the lower part of the face. This includes the mouth region and to a lesser extent, the nose region, which are not highly informative about gender for neutral-expression faces (see Figure 5a). In the implementation of this concept within the context of an FIO model, we use an altered internal representation of the face stimuli. As seen in Figure 6d, the internal face representations are altered by lowering the contrast of the mouth region by 80%, and the contrast of the nose region by 60%, which results in a diminished representation of those regions. Aside from this difference in internal representation, the FIO model remains unchanged. As with previous iterations of the model, the internal noise was fit to the human data from neutral-expression stimuli. However, for this model, we did a damping parameter search across 125 combinations of three additional contrast parameters for the eye region, nose region, and mouth region. The parameter combinations were uniformly distributed across different values of contrast for each region, so that we also tested combinations where the eyes were dampened instead of the mouth and combinations where all features were dampened similarly. These parameters were fit to minimize the combined AIC for happy-expression and neutral-expression human data compared to the model. The results (Figure 6e) show that the model with a highly dampened representation of the mouth region, and a somewhat dampened representation of the nose

region, is able to better capture important aspects of the human performance curve. First, the model's performance with happy-expression faces is now comparable to that with neutral-expression faces, in agreement with the human results. In addition, the fit of this FIO model to the human forced fixation performance is improved compared to the previous models (AIC of .375 for neutral-expression faces and .16 for happy-expression faces).

CNN model separately trained on happy and neutral faces supports the results of the IO and ROI models.

In addition to running altered versions of the FIO, we also run a CNN model that is able to simulate the development of an internal face representation in humans based on exposure to a training set with real-world statistics of happy-expression and neutral-expression faces. We use the statistics of facial expressions that we found with the eyewear embedded camera and train a CNN model on a gender discrimination task with a stimulus set of 89% neutral-expression faces and then 11% happy-expression faces. Then we test each of the models with a separate set of neutral-expression faces and a separate set of happy-expression faces, which are the same stimuli that were used in human psychophysics experiments above. Figure 7a shows a visualization of the most important parts of the stimuli that the model uses to do the classification task. The visualizations of the model tested on happy-expression faces and neutral-expression faces look similar, with a large use of the eye region but not the mouth region. This suggests that a model trained mostly on neutral-expression faces is unable to learn to use the extra gender information present in happy-expression faces during testing. This is in contrast to a CNN model that was trained on only happy-expression faces (Figure 5c), that is able to use this extra information, which suggests

that these two models have different internal representations of faces. Figure 7b shows the corresponding performance of the model trained on mostly neutral-expression faces when it is separately tested on neutral-expression faces and happy-expression faces. Performance is highest when it is tested on neutral-expression faces, which suggests that the internal representations of the CNN model are sensitive to the statistical frequency of facial expressions that are used during the training phase.

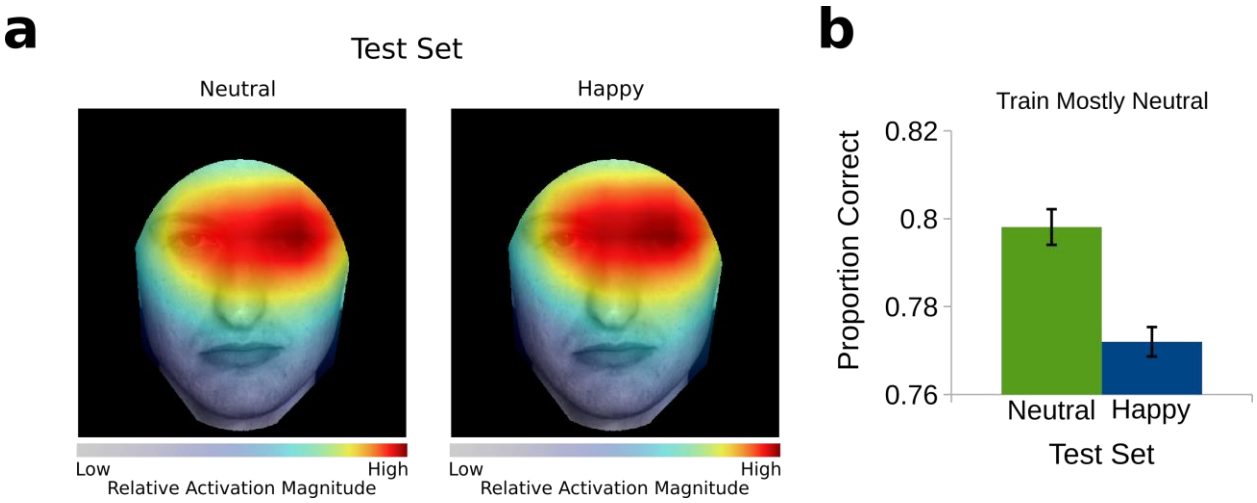

Figure 7: ***a)*** *Visualizations are shown of the parts of the face stimuli that the CNN model uses the most when it is trained on a real-world frequency of happy-expression and neutral-expression faces in a face gender discrimination task and then separately tested on neutral-expression faces and happy-expression faces, respectively. The activation magnitude scale used to show the importance of different face features is relative only within each image because each visualization has been normalized such that the features with the highest activations are mapped to the highest values. The visualizations show that the eyes are an important region that the CNN uses to do the classification task, regardless of which stimuli were used for testing. Even though the CNN is trained on some happy-expression stimuli, it does not learn to use more information from the mouth region because the frequency of happy-expression stimuli during training is very low.* ***b)*** *Corresponding performance is shown for each of the two testing sets used to run this CNN model. Performance is highest when the model is tested on neutral-expression faces, relative to when it is tested on happy-expression faces. The Error bars represent the standard error of the mean of 20 training runs for each model.*

3.4 Discussion

Previous studies have shown that the initial human eye movement to a face is critical and sufficient to achieve close to 90% of asymptotic performance in face discrimination tasks (Hsiao and Cottrell, 2008; Or et al., 2015). In addition, studies have shown that the brain plans initial eyes movements to locations within faces that maximize the acquisition of information to support evolutionarily important tasks (Or et al., 2015; M. F. Peterson & Eckstein, 2012).

Here, we asked whether the human initial fixation strategy for a gender discrimination task is fixed or can adapt to specific distributions of information in faces. In laboratory search tasks, studies have shown varying degrees of adaptability of eye movements with changes in the prior probability or rewards associated with various stimuli and locations (Ackermann & Landy, 2010; Droll, Abbey, & Eckstein, 2009; M. P. Eckstein, Schoonveld, Zhang, Mack, & Akbas, 2015b; Liston & Stone, 2008; Navalpakkam, Koch, Rangel, & Perona, 2010; M. S. Peterson & Kramer, 2001; Stritzke, Trommershäuser, & Gegenfurtner, 2009; Walthew & Gilchrist, 2006). Unlike the flexible oculomotor plans in simpler synthetic tasks, we found that eye movement plans are less flexible to variations in the distribution of gender information across faces with different emotional expressions.

We showed that humans, unlike ideal observers and foveated ideal observers, are unable to take full advantage of gender information in the mouths of happy-expression faces, related to differences in muscle activation during smiling (Soussignan et al., 2013). The increased gender information in the mouth region for happy faces leads to a downward shift in the theoretical optimal initial point of fixation. However, the empirically measured human optimal point remains unchanged, and humans show only a slight downward shift in their

initial fixation position during free-viewing. These findings are consistent with evidence that the information in the mouth region in general is only used by humans at a course spatial scale for gender discrimination (Schyns et al., 2002; Smith, Cottrell, Gosselin, & Schyns, 2005). As a result, details that are represented by fine spatial frequencies such as teeth and corners of the mouth may be largely left unused. One possible explanation is that humans do not use much information from of the mouth region in general during face discrimination tasks, rather than only in gender discrimination with happy-expression faces. However there is evidence against this interpretation since (M. F. Peterson & Eckstein, 2012) showed that humans shifted their initial eye movement downward in a happy-expression vs. neutral-expression discrimination task to take advantage of the crucial information in the mouth region. In addition (Schyns, Bonnar, & Gosselin, 2002) showed that observers use fine spatial frequency details in the mouth region both for face identification and expressive vs. neutral-expression face discrimination tasks.

Here, we showed that there are scenarios, such as with happy-expression faces, where the human strategy to not utilize gender information in the mouth area is suboptimal. Why would humans show such suboptimal strategy? We hypothesized that our findings might reflect a strategy by humans that is adapted to the statistical occurrence of facial expressions in the natural world. If neutral expressions were more prevalent than happy ones, then we might expect the human eye movement and feature utilization strategy to be more adapted to faces with neutral expressions. Our findings are consistent with this interpretation. First, our measurements of the naturally occurring emotional expressions encountered by subjects with an eye-wear-mounted camera suggested a vastly more frequent appearance of neutral expressions compared to all other expression categories. It is possible that the expression

frequency measurements were partly influenced by the encountered people who realized that the participants were wearing an unusual eye-wear-mounted camera. However, if anything, we might expect that such a scenario would only increase the frequency of non-neutral expressions such as surprise and happiness. As a result, our measured frequency might be a lower limit of the frequency of neutral expressions. Second, our psychophysical findings suggest a strategy that agrees quite well with the Foveated Ideal Observer for neutral face expressions. Third, our results with simulating differences in human development of face information extraction with a CNN model trained on neutral-expression stimuli, and separately trained on happy-expression stimuli, support our ideal observer analysis. We show that a model trained on neutral-expression faces and tested on happy-expression faces has lower performance and is unable to make use of additional gender information in the mouth region relative to a model trained and tested on happy-expression faces. Although they are only a rudimentary approximation of human cortical processing, CNNs are starting to be used in the study of human vision and face processing (see (O'Toole, Castillo, Parde, Hill, & Chellappa, 2018) for a review) after successful implementations of various face classification tasks in computer vision (Li, Lin, Shen, Brandt, & Hua, 2015; Schroff, Kalenichenko, & Philbin, 2015; Taigman, Yang, Ranzato, & Wolf, 2014) , some of which have achieved close to human performance. CNNs are known to have certain useful properties that may be able to represent aspects of the human visual system. One of those aspects is a feedforward multilayer structure that represents progressively more complex features starting from edge detection and ending with complex shapes, textures, colors, and the relationships between them. Another important aspect is the ability to learn feature detectors that are adapted to the complex statistical properties of the features in the images that the model is being trained on.

In the context of the gender discrimination task presented here, along with the data on the frequency of human facial expressions, humans are overwhelmingly exposed to neutral-expression face stimuli, rather than happy-face stimuli. We are able to represent these differences by training a CNN model on datasets with different statistics of facial expressions and showing corresponding differences in performance as well differences in the location of important information that is being used for the task.

When taking the foveation of the visual system into account, we attempted to use an FIO model that incorporated the frequency of neutral and happy expressions in humans during everyday interactions, as prior probabilities. However, this model did not fit the human data well, and was very sensitive to changes in the prior probabilities. A model that used a diminished representation of the mouth provided the best fit to human data instead. It is important to note the difference between changing the priors of facial expressions vs. changing representations of the faces themselves and that the neural correlates of these manipulations may relate to different stages of face processing (see (Tsao & Livingstone, 2008) for a review). Changing the priors is a manipulation that affects a decision variable that is formulated in late stages of processing. In humans this would correspond to processing the same face stimuli in a similar way for different tasks, until a late stage where there is a top-down mechanism that affects a perceptual decision based on task demands. This mechanism would affect the perceptual decision by using prior perceptual experiences encoded in memory. For example, in this view, humans would process faces in a discrimination task between happy and neutral faces in the same way as in a gender identification task with happy faces, until a late decision-making stage that involves a low dimensional decision variable. In contrast, changing the representation of the mouth in the model would

correspond to an earlier effect of top-down task specific mechanisms. In this view, the same face stimuli would be processed differently, earlier in the face processing stream, depending on the current task. For a gender identification task, this model represents a mechanism where the more important parts of the face (eyes and nose) are extracted from the original face stimulus for further processing.

In the larger context of vision science, our findings can be related to the increasing evidence of how the visual system is tuned to statistical regularities in the environment. This happens at multiple levels in the visual stream, including at a low-level of spatiotemporal sensitivity in the retina, in early cortical areas, in mid-level vision (Burge & Geisler, 2014, 2015; Burge & Jaini, 2017; Geisler & Ringach, 2009; S. Zhang, Abbey, & Eckstein, 2009), as well as at higher level of object-recognition and use of scene context (Bar, 2004; M. Eckstein, 2017; M. P. Eckstein, Koehler, Welbourne, & Akbas, 2017; Hidalgo-Sotelo, Oliva, & Torralba, 2005; Koehler & Eckstein, 2017; Torralba, Oliva, Castelhano, & Henderson, 2006). This tuning is thought to be driven by a combination of evolutionary development and direct experience with visual stimuli in the environment. Faces are thought to be the most complex of objects that humans are frequently exposed to and it is known that there are specialized areas of the brain responsible for specific aspects of face perception (Dachille, Gold, & James, 2012; Haxby, Hoffman, & Gobbini, 2000; Kanwisher, McDermott, & Chun, 1997). A well-known result in face perception research is that performance in face discrimination tasks drops substantially when faces are inverted (upside-down) relative to upright faces (Farah et al., 1995; Sekuler et al., 2004). It is thought that due to the very infrequent prevalence of inverted faces in the experience of observers, they are not processed with the same efficiency by face-specific brain areas despite the fact that all of the lower

level properties of the stimuli themselves are the same as in upright faces. Given the specificity of face processing, it may be that information for specific tasks is extracted most efficiently for the most commonly seen facial expressions during those tasks. Our findings expand these results to eye movements and facial feature utilization for gender discrimination, suggesting that humans adopt a strategy that optimizes gender discrimination to the emotional expressions most frequently encountered in everyday life.

4 The development of internal fixation-specific face representations

4.1 Introduction

Face perception is an important ability that most people use many times a day in the context of various common tasks such as face identification, gender discrimination, and emotion discrimination. There is a large amount of research on the specificity of face processing in the brain in relation to how it differs from the processing of other complex objects (references). There are many aspects of this specificity, including, the specialization of face-processing areas in the brain, the encoding of task-specific information…., and the extraction of information from faces with eye movements to specific areas on the face. Here, we focus on the last aspect and how it may relate to individual-specific eye movement strategies to faces.

There is a diminished quality of visual processing by areas of the retina outside the foveal region, due to a lower density of photoreceptors. Eye movements, which point the high acuity foveola region of the retina during exploration of the environment, are critical for accomplishing evolutionarily important tasks. The brain programs eye movements by taking into account the foveated properties of the visual system in conjunction with the distribution of task-relevant information in the environment (G. E. Legge et al., 1997; Gordon E. Legge et al., 2002; Najemnik & Geisler, 2005, 2009; Paulun et al., 2015; M. F. Peterson & Eckstein, 2012) to maximize the acquisition of information during basic

perceptual tasks (optimal or near-optimal oculomotor strategies).

During face discrimination tasks, humans exhibit high levels of accuracy with even just a single eye-movement (Hsiao & Cottrell, 2008; Or et al., 2015). It has been shown that there exists an empirically optimal point of initial fixation to a face located below the eyes that for most observers leads to the highest performance in various face discrimination tasks (M. F. Peterson & Eckstein, 2012). This fixation location is also theoretically optimal, as predicted by a computational model that incorporates a representation of the fovea, but otherwise makes optimal decisions under uncertainty. Furthermore, it has been shown that this initial optimal point is very consistent across observers, as well as across time within the same observer. In addition, this result has been reproduced outside of laboratory conditions with participants viewing faces while walking around in the real world with mobile eyetrackers (M. F. Peterson et al., 2016). In face discrimination tasks, an observer's ability to make an eye movement to their empirical optimal point of fixation is an important determinant of their ability to maximize their performance in that task.

Despite a strong consistency of the location of the initial fixation to faces across a large portion of observers to a position slightly below the eyes, there is a small percentage of observers (about 10%) that consistently fixate a position lower on the face, around the tip of the nose, and even more rarely around the mouth (M. F. Peterson & Eckstein, 2013). There is a distribution of vertical fixation positions to faces across observers, with the vast majority of initial fixations located around the eye region, with a long tail down toward the nose and mouth region. For simplicity, we call those that fixate closer to the eyes, "eye-lookers", and those that fixate closer to the nose, "nose-lookers." The (M. F. Peterson & Eckstein, 2013) study showed that the fixation location that is theoretically optimal, as shown with a

Foveated Ideal Observer (FIO) model, and empirically optimal for eye-lookers, as measured in performance at various fixation locations, was actually suboptimal for nose-lookers. Furthermore, the lower preferred point of initial fixation was shown to be empirically optimal for nose-lookers. However, it is unknown what mechanism causes these individual differences in observers.

Here we build on the previous research in individual differences of the initial eye movement to faces and test two different theories to try to explain this effect. The first theory, which we refer to as the "altered-anisotropy theory," involves possible differences in the anisotropy of the retina between eye-lookers and nose-lookers. It is known that the human retina has differences in the density of photoreceptors as well as retinal ganglion cells at different locations that correspond to different parts of the visual field. Besides a general loss of spatial acuity with increasing eccentricity from the center of the visual field (Golla, Ignashchenkova, Haarmeier, & Thier, 2004; Yeshurun & Carrasco, 1999), differences in the quality of representation of different parts of the visual environment are also known to exist both between the lower and upper visual field (vertical anisotropy) (Marisa Carrasco, Talgar, & Cameron, 2001; Corbett & Carrasco, 2011), as well as between the vertical and horizontal directions (horizontal-vertical anisotropy) (Cameron, Tai, & Carrasco, 2002; MARISA Carrasco & Frieder, 1997). Here we focus on the vertical anisotropy between the upper and lower visual field, which has been shown to result in higher performance when stimuli are presented in the lower visual field relative to when the same stimuli are presented in the upper visual field in both simple low-level visual tasks (Marisa Carrasco et al., 2001; Corbett & Carrasco, 2011) as well as in more complex higher-level (Marisa Carrasco, Marie Giordano, & McElree, 2004; S. He, Cavanagh, & Intriligator, 1996; Intriligator & Cavanagh,

2001; Kristjánsson & Sigurdardottir, 2008) visual tasks. In relation to faces, individual-specific differences in vertical anisotropy may result in different vertical fixation behaviors to faces. More specifically, we investigate if nose-lookers have a ratio of acuity in their upper visual field relative to their lower visual field that is higher than the same ratio in eye-lookers. For example, if nose-lookers have higher acuity in the upper visual field compared to eye-lookers, then nose-lookers may look lower on the face than eye-lookers because they are able to get a similar quality of input from the eyes even if they look further down.

The second theory that we explore, which we refer to as the "matched-template" theory, does not involve differences in low-level visual processing that can be generalized to non-face tasks. Instead, this theory involves differences between eye-lookers and nose-lookers that may be explained by a difference in the representation of a face template that is stored in higher-order brain areas. There is reason to believe that fixation-specific tuning may be present in human face-selective neurons based on the findings of several studies in humans and monkeys. There is evidence for increased neural activity in the posterior lateral face patch in monkeys (Issa & DiCarlo, 2012) and increased separation of activity patterns to major face features in the right inferior occipital gyrus in humans (de Haas et al., 2016), when the facial features appear at the typical retinal positions relative to a preferred point of fixation, compared to atypical locations. In addition, a "retinotopic protomap" has been proposed as an organizing principle for higher visual areas, such as the FFA and OFA (Hasson, Levy, Behrmann, Hendler, & Malach, 2002), where the exposure of the visual system to consistent locations of objects in certain ranges of eccentricity creates a bias such that neurons that are tuned to those objects are preferentially activated when those objects are presented at specific locations in the visual field. In relation to this theory, a more recent

study also found evidence of "faciotopy," which refers to a cortical map that topographically represents features of a face itself, rather than just representing a part of the observers visual system, such as with retinotopy (Henriksson, Mur, & Kriegeskorte, 2015). In this study, it was found that the distances between the face-feature-selective patches of cortex reflected the physical distances between the actual features in a face stimulus. If representations of faces are organized in a fixation-specific way in the brain, then it may be possible that there are individual differences in those representations that are shaped by a long-term initial-fixation strategy to a specific part of the face.

Here, we measure human eye movements and performance profiles (performance at different forced-fixation positions down the vertical midline of a stimulus) at different fixation positions during various tasks and compare results between groups of eye-lookers and nose-lookers. We measure the location of the initial eye movement during several different face discrimination tasks, including human identification, famous face identification, emotion identification, and gender identification to determine fixation consistency across these tasks and separate participants into eye-lookers and nose-lookers. We also measure the location of the initial eye movement in a chimp face identification task, a luggage bag identification task, and a sports/regular car identification task to compare eye movement behavior between human-face and non-human-face tasks. We then measure performance profiles for eye-lookers and nose-lookers during a human face identification task, a sports/regular car identification task, a gabor detection task, and a natural image matching task. We run the first three tasks in order to measure differences in the location of optimal (performance-maximizing) fixation positions between eye-lookers and nose-lookers and determine if those differences correlate with possible differences in vertical anisotropy

between the two participant groups. The sports/regular car identification task is run to see if the optimal fixation behavior that is observed with face stimuli can generalize to a non-face task. In addition to running experiments with human participants, we run two computational models to see if we can reproduce differences between eye-lookers and nose-lookers based on differences in a face template that is presented to the models. The first model is a Foveated Ideal Observer (FIO), which simulates the foveation of the visual system, but otherwise makes optimal decisions under uncertainty (pixel noise) in any classification task that it is run on. The second model is a convolutional neural network (CNN), which is suboptimal, but is used to represent a rudimentary version of human development with different face templates.

The analysis of the human eye movement tasks shows that there is a strong correlation in vertical fixation position between tasks that involve human faces, with a lower correlation to a chimp face discrimination task, and even lower correlation to a luggage bag discrimination task. The analysis of performance in the forced fixation tasks, comparing the human face identification task with the gabor detection and natural image matching task, shows there are no significant differences in vertical anisotropy between eye-lookers and nose-lookers, which provides evidence against the altered-anisotropy theory in favor of the matched-template theory. In addition, comparison of the forced-fixation human face identification task with the forced-fixation gender identification task and the forced-fixation sports/regular car identification task shows that the specificity of eye movements to human faces, as represented by the FIO, generalizes between human face tasks, but does not generalize to other stimuli. Finally, in support of the matched-template theory, the results of a modified FIO with fixed face templates (FT-FIO) along with CNN simulations show that

forced-fixation performance profile differences between eye-lookers and nose-lookers can be qualitatively reproduced by representing differences in a representation of a fixation-dependent face template.

4.2 Materials and Methods

Human Psychophysics Studies.

Many different tasks are presented in this paper, some of which were done several years apart and have used different monitor setups. For conceptual clarity, we describe the tasks in the order that they are presented in the results section of the paper rather than the order in which they were actually completed. For most of the methods sections below, we group the completed experiments into three groups with multiple tasks in each group and refer to them in the description of each section:

Experiment Group 1.

This experiment group consisted of 7 different free-fixation tasks where participants were free to make eye movements to the presented stimuli.

Task 1. Human Face Identification with 10 Caucasian male faces: Free-Fixation

Task 2. Human Face Gender Discrimination: Free-Fixation

Task 3. Human Famous Faces Identification: Free-Fixation

Task 4. Human Emotion Discrimination: Free-Fixation

Task 5. Human Face Identification with 4 Caucasian male faces: Free-Fixation

Task 6. Chimp Face Identification: Free-Fixation

Task 7. Luggage Bag Identification: Free-Fixation

Experiment Group 2.

This experiment group consisted of 2 different free-fixation tasks where participants were free to make eye movements to the presented stimuli, as well as 2 different forced-fixation tasks where participants were forced to fixate certain points along the presented stimulus during the duration of a trial.

Task 1. Human Gender Discrimination: Free-Fixation

Task 2. Human Gender Discrimination: Forced-Fixation

Task 3. Regular/Sports Car Discrimination: Free-Fixation

Task 4. Regular/Sports Car Discrimination: Forced-Fixation

Experiment Group 3.

This experiment group consisted of 3 different forced-fixation tasks where participants were forced to fixate certain points along the presented stimulus during the duration of a trial.

Task 1. Human Face Identification with 10 Caucasian male faces: Forced-Fixation

Task 2. Natural Scenes Matching: Forced-Fixation

Task 3. Single Gabor Detection: Forced-Fixation

Participants.

Set 1. A group of 25 undergraduate students of either sex participated for course credit. They completed Tasks 1-4 in Experiment Group 1, as well as all Tasks in Experiment Group 3.

Set 2. A group of 78 undergraduate students of either sex participated for course credit. They completed Tasks 5-7 in Experiment Group 1.

Set 3. A group of 6 undergraduate students of either sex, participated for course credit. They completed all Tasks in Experiment Group 2.

Informed consent was obtained from all subjects and guidelines provided by the institutional review board of the University of California, Santa Barbara were followed.

Apparatus and Materials.

Setup 1.

This setup was used for Tasks 1-4 in Experiment Group 1, as well as all Tasks in Experiment Group 3. MATLAB Psychtoolbox and Eyelinktoolbox software were used to run the eyetracker from a display computer as well as present visual stimuli on the display screen. The display was a CRT monitor set to a 800x600 pixel resolution and was located 46.5cm away from the observer's eyes. The display was linearly calibrated with a minimum luminance of .05 cd/m² and a maximum luminance of 50 cd/m².

Setup 2.

This setup was used for Tasks 5-7 in Experiment Group 1 as well as for all Tasks in Experiment Group 2. MATLAB Psychtoolbox and Eyelinktoolbox software were used to run the eyetracker from a display computer as well as present visual stimuli on the display screen. The display used was a Barco MDRC 1119 monitor set to a 1280x1024 pixel resolution and was located 76.5cm away from the observer's eyes. The display was linearly calibrated with a minimum luminance of .05 cd/m² and a maximum luminance of 126 cd/m².

Eye-tracking. The same eyetracker settings were used for all experiment groups and tasks. The left eye of each participant was tracked using an SR Research Eyelink 1000 Tower Mount eye tracker sampling at 250 Hz. A nine-point calibration and validation were run before each 125-trial session, with a mean error of no more than 0.5º of visual angle. Saccades were classified as events in which eye velocity was greater than 35º and eye acceleration exceeded 9,500º per square second. The recommended thresholds by SR for cognitive research are an eye velocity of 30º and an eye acceleration of 8,000º per square second. The minor increase of the velocity and acceleration thresholds in our parameter settings allowed us to better control the number of "broken fixations" during the initial

fixation stage at the beginning of every trial prior to the presentation of the stimulus.

Stimuli.

Experiment Group 1.

Human Face Identification Task with 10 Faces: As seen in Figure 1a, the stimuli for this task were used for both a free-fixation and forced-fixation condition (in Experiment Group 3). The stimuli consisted of 10 grayscale front-view Caucasian male faces. All of the images were taken with constant diffuse lighting, distance, and camera settings. A Canon digital camera was used. The digital pixel value was a non-linear saturating function of luminance (a standard Canon log-cine transfer function). The images were normalized by scaling and cropping, such that the center of the eyes was 2/5 of the image height below the top of the image and the chin was 1/50 of the image height above the bottom of the image. The faces were luminance-mean normalized to 25 cd/m² and shown to participants at a Root Mean Square (RMS) contrast of 0.13, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 1.97 cd/m² (corresponding to a noise RMS contrast of 0.079). Participants viewed the face stimuli 46.5cm away from the display resulting in a square stimulus (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. The large size of the faces, more typical of conversational distance, was chosen: (1) to allow measurements of larger variations of perceptual performance with fixation position (for small faces perceptual performance is less sensitive to fixation position within the face); (2) to allow more precise measurements of fixation positions relative to facial features. In addition, the large faces (e.g., 10 º width, 15 º height) have been shown to optimize face identification (Yang et al., 2014).

Human Gender Discrimination Task: As seen in Figure 1b, the stimuli for this task were used only in a free-fixation condition. The stimuli consisted of 80 (40 male and 40 female) grayscale front-view Caucasian faces. All of the images were taken using the same camera and settings as in the *Human Face Identification Task* above. They were also spatially aligned and presented on screen with the same the same size and mean luminance level as in the *Human Face Identification Task* above. However, an RMS contrast of 0.226 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 2.95 cd/m² (corresponding to a noise RMS contrast of 0.118).

Human Emotion Discrimination Task: As seen in Figure 1c, the stimuli for this task were used only in a free-fixation condition. The stimuli consisted of 140 (70 male and 70 female) grayscale front-view Caucasian faces with 7 emotions (10 images of each gender for each emotion): neutral, happy, sad, afraid, surprised, angry, and disgusted. All of the images were taken using the same camera and settings as in the *Human Face Identification Task* above. They were also spatially aligned and presented on screen with the same the same size and mean luminance level as in the *Human Face Identification Task* above. However, an RMS contrast of 0.224 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 2.95 cd/m² (corresponding to a noise RMS contrast of 0.118).

Human Famous Face Identification Task: As seen in Figure 1d, the stimuli for this task were used only in a free-fixation condition. The stimuli consisted of 100 color front-view Caucasian faces of either gender. All of the images were found from various sources of celebrity photos on the internet. They were also spatially aligned and presented on screen with the same the same size as in the *Human Face Identification Task* above. However, they

were shown in color, with full contrast and no noise, since this task did not measure performance.

Human Face Identification Task with 4 Faces: As seen in Figure 1e, the stimuli for this task were used only in a free-fixation condition and are a subset of 4 of the 10 spatially aligned face images used in the *Human Face Identification Task* of Experiment Group 1. Only 4 images were used in order to make it easier for participants to quickly learn the faces and complete a short task. participants viewed the face stimuli 76.5cm away from the display resulting in square stimuli (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. They were luminance mean normalized to 64 cd/m². An RMS contrast of 0.152 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 7.06 cd/m² (corresponding to a noise RMS contrast of 0.11).

Chimp Face Identification Task: As seen in Figure 1f, the stimuli for this task were used only in a free-fixation condition and consisted of 4 front-view male chimp faces. These images were taken from a larger set of front-view chimpanzee faces obtained from the lab of Lisa A. Parr and are similar to the front-view chimp faces seen in (Parr, Heintz, Lonsdorf, & Wroblewski, 2010). Only 4 images were used in order to make it easier for participants to quickly learn the chimp faces and complete a short task. The chimp faces were spatially aligned, using the position of the eyes and chin, the same way as with the stimuli in the *Human Face Identification Task* above. Participants viewed the face stimuli 76.5cm away from the display resulting in square stimuli (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. They were luminance mean normalized to 64 cd/m². An RMS contrast of 0.173 was used, where part of that

contrast variation came from added Gaussian white noise with a standard deviation of 7.06 cd/m² (corresponding to a noise RMS contrast of 0.11).

Luggage Bag Identification Task: As seen in Figure 1g, the stimuli for this task were used only in a free-fixation condition and consisted of 4 luggage bags. The background body of each of the 4 bags was the same image, but the bags were manipulated in Photoshop so that they differed only in three features along the body of the bag: a logo at the top, a logo in the middle, and the position of the wheels at the bottom. Each bag had a unique version of each of the three features, such that any of the features could be used to distinguish the bags. The bags were spatially aligned, vertically and horizontally. Participants viewed the face stimuli 76.5cm away from the display resulting in square stimuli that subtended 18 ° in width and height. They were luminance mean normalized to 64 cd/m². An RMS contrast of 0.22 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 7.06 cd/m² (corresponding to a noise RMS contrast of 0.11).

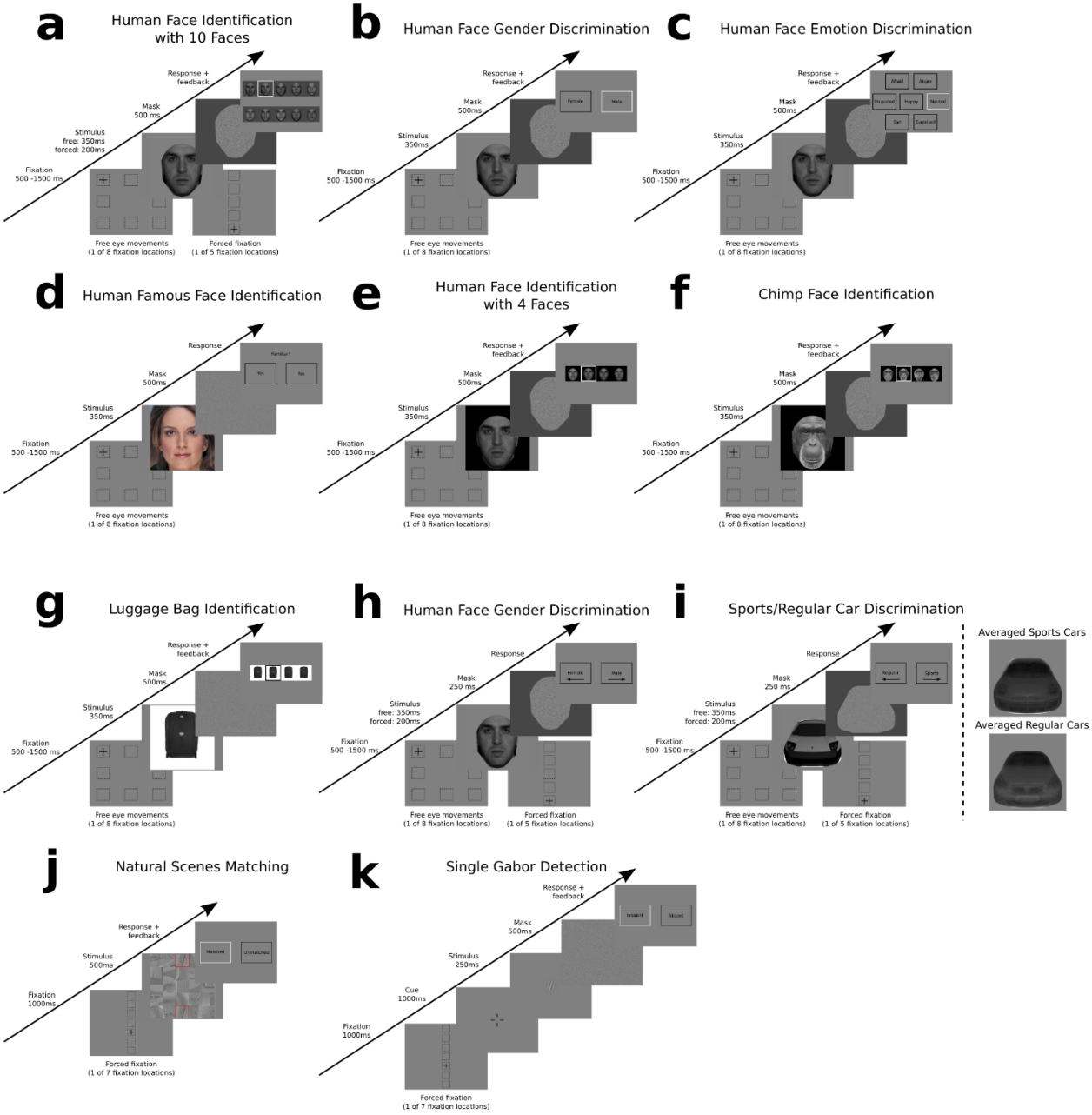

*Figure 6: (**a**) A trial timeline for a human face identification task with 10 faces is shown. This task had a free-fixation component (bottom left panel) and a forced-fixation component (bottom right panel). In the free-fixation component, participants started with a fixation at one of 8 points on the periphery of the screen and were then free to make eye movements to the stimulus when it was shown. In the forced-fixation component they were forced to keep their fixation at specific points on the vertical midline of the face corresponding to the forehead, eyes, nose, mouth, and an individual preferred point found during the free-fixation condition. The stimulus was followed by a noise mask to limit visual and cognitive aftereffects. Participants chose 1 of 10 faces with a mouse on a response screen at the end of the trial and immediately received feedback.*

*(**b**) A trial timeline of a free-fixation human face gender discrimination task is shown. At the end of a trial, participants selected with the mouse one of two genders written on the screen and then immediately received feedback.*

*(**c**) A trial timeline of a free-fixation human famous face emotion discrimination task is shown. At the end of a trial, participants selected with the mouse one of seven emotion written on the screen and then immediately received feedback.*

*(**d**) A trial timeline of a free-fixation human famous face identification task is shown. At the end of a trial, participants selected with the mouse either "yes" or "no" to answer whether they are familiar with the face being shown. No feedback was given since there was no correct answer.*

*(**e**) A trial timeline of a free-fixation human face identification task with 4 faces is shown. The trial timeline is the same as for the free-fixation component shown in (a), except that only a subset of 4 of the original 10 faces were used in order for participants to be able to achieve a reasonable level of performance in a single sitting. At the end of a trial, participants selected with the mouse one of four faces shown on the screen and then immediately received feedback.*

*(**f**) A trial timeline of a free-fixation chimp face identification task is shown. The trial timeline is the same as in (e), except that 4 chimps faces were used instead of 4 human faces.*

*(**g**) A trial timeline of a free-fixation luggage bag identification task is shown. The trial timeline is the same as in (e), except that 4 images of luggage bags were used instead of 4 human faces.*

*(**h**) A trial timeline of a human face gender discrimination task is shown with a free-fixation (bottom left panel) and forced-fixation (bottom right panel) component. In the forced-fixation component, participants were forced to keep their fixation at specific points on the vertical midline of the face corresponding to the forehead, eyes, nose, mouth, and chin. At the end of a trial, participants selected with a right arrow left arrow keyboard press, one of two genders written on the screen. Arrows were also drawn on the response screen to remind participants which keyboard press corresponded to which gender. Participants did not receive feedback.*

*(**i**) A trial timeline of a sports/regular discrimination task is shown with a free-fixation (bottom left panel) and forced-fixation (bottom right panel) component. In the forced-fixation component, participants were forced to keep their fixation at one of 5 specific points on the vertical midline of the car equally spaced from the roof down to the license plate. At the end of a trial, participants selected with a right arrow left arrow keyboard press, one of two car categories (sports or regular) written on the screen. Arrows were also drawn on the response screen to remind participants which keyboard press corresponded to car category. Two images on the right side show averages of the car images from each category. Participants did not receive feedback.*

*(**j**) A trial timeline of a forced-fixation natural scenes matching task is shown. A trial began with a forced fixation at 1 of 7 specific points on the vertical midline of the screen equally spaced from the top to the bottom. A 5x5 grid of small squares with different natural scenes was then shown in the center of the screen. There was a 50% probability on each trial that the scene in the square at the top center of the grid was the same one as the scene in the*

square at the bottom center of the grid. At the end of a trial, participants selected with the mouse whether the top and bottom squares in the vertical center of the grid were "matched" or "unmatched" and then immediately received feedback.

***(k)** A trial timeline of a forced-fixation single gabor detection task is shown. A trial began with a forced fixation at 1 of 5 specific points on the vertical midline of the screen equally spaced from the top to the bottom. A cue was then briefly shown in the center of the screen to guide the participants' attention to that location. A single gabor was then shown in the center of the screen with a 50% probability on each trial at the location of the cue. A noise mask followed after that to limit visual and cognitive aftereffects. At the end of a trial, participants selected with the mouse whether a gabor was "present" or "absent" in the center of the screen during the trial and then immediately received feedback.*

Experiment Group 2.

Human Gender Discrimination Task: As seen in Figure 1h, the stimuli for this task were used in both a free-fixation and a forced-fixation condition. The stimuli are the same images and were spatially aligned in the same way as in *Human Gender Discrimination Task* of Experiment Group 1. However, in this experiment a different setup was used where participants viewed the face stimuli 76.5cm away from the display resulting in square stimuli (face and mask) that subtended 18 º (~15 º for the part of the face that is not covered with the mask) in width and height. They were luminance mean normalized to 25 cd/m². An RMS contrast of 0.151 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 2.75 cd/m² (corresponding to a noise RMS contrast of 0.11).

Car Discrimination Task: As seen in Figure 1i, in this task, participants identified if a given car was a sports car or a regular car. The stimuli consisted of 20 regular cars and 20 sports cars. The stimulus set was created by rendering frontal photographs of 3D models of cars available freely online from websites like TurboSquid and CGTrader. Using Autodesk 3DsMax, these 3D models were placed in a standard scene with diffuse sky lighting. A camera was placed at a consistent distance, offset vertically above the mid plane of the car

pitching downwards by about 6°. This angle was necessary to occlude the wheels which are often very distinctive for sports cars. The right side of Figure 1i shows an average (mean of corresponding pixel values across the images) of all the regular car images compared to the average of all the sports car images. During the discrimination, a standardized mask was used to crop the images such that only internal features of the cars (head lamps, radiators, wind shield, hood and side mirrors) could be used for discriminating the cars. Participants viewed the car stimuli 76.5cm away from the display resulting in square stimuli (car and mask) that subtended 18 º (~15 º for the part of the car that is not covered with the mask) in width and height. They were luminance mean normalized to 25 cd/m². An RMS contrast of 0.266 was used, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 2.75 cd/m² (corresponding to a noise RMS contrast of 0.11).

Experiment Group 3.

Human Face Identification Task with 10 Faces: This was a forced-fixation task. The same stimuli were used for this task as in the free-fixation Human Face Identification Task with 10 Faces described in Experiment Group 1, as seen in Figure 1a.

Natural Images Matching Task: As seen in Figure 1j, the stimuli for this task were used only in a forced-fixation condition. The stimulus consisted of a grid of 5x5 grid of small square images with a width and height of 10.84 º . The grid contained a random sample of 25 images of natural scenes taken from a larger set of 1000 images. 2000 grids of 25 images were premade for quick stimulus presentation during the task, where 1000 of them contained the same “matched” image in the middle of the top row and the middle of the bottom row (see the stimulus example in the center box of Figure 1j), and the other 1000 did not contain a match. During presentation, the grids were luminance mean normalized to 25 cd/m² and

shown to participants at an RMS contrast of 0.125, where part of that contrast variation came from added Gaussian white noise with a standard deviation of 0.984 cd/m² (corresponding to a noise RMS contrast of 0.0394).

Single Gabor Detection Task: As seen in Figure 1k, the stimuli for this task were used only in a forced-fixation condition. The stimuli consisted of a gabor patch with a frequency of 4 cycles/degree and a Gaussian envelope with a standard deviation of .75 º. The stimulus was mean luminance normalized to 25 cd/m² and shown to participants at a contrast level was determined individually for each participant using a staircase procedure set to a performance (percent correct) threshold of 85%. An additional contrast variation came from added Gaussian white noise with a standard deviation of 4.92 cd/m² (corresponding to a noise RMS contrast of 0.197).

Procedure.

The tasks described below are separated into different experiment groups for conceptual clarity and consistency with the Results section. However, Tasks 1-4 in Experiment Group 1 were completed by the same participants as all Tasks in Experiment Group 3. The combination of those tasks was completed in a random order for each participant. Tasks 5-7 in Experiment Group 1 were completed by a separate group of participants, each of whom did them in random order. In Experiment Group 2, the tasks were completed with a separate group of participants in alternating blocks with the *Human Gender Discrimination Task* assigned to odd-numbered blocks and the *Car Discrimination Task* assigned to even-numbered blocks.

Tasks 5-7 in Experiment Group 1 were completed by each participant in a single sitting of no longer than 2 hours. For all other experiment groups and tasks, participants

completed the tasks in multiple eyetracking sessions, each of which lasted no longer than 2 hours.

Experiment Group 1.

Human Face Identification Task with 10 Faces: This task consisted of a free-fixation component which had to always be run before a forced-fixation component in order to obtain an observer-specific preferred fixation position, which was then used as one of the 5 forced-fixation points in the forced-fixation component. The free-fixation component consisted of 2 blocks, with 100 trials per block. This was followed by forced-fixation component (in Experiment Group 3), which consisted of 1 practice block with 12 trials, followed by 7 experimental blocks with 150 trials each.

Human Gender Discrimination Task: This task contained only a free-fixation component and consisted of 2 blocks, with 100 trials per block.

Human Emotion Discrimination Task: This task contained only a free-fixation component and consisted of 2 blocks, with 100 trials per block.

Human Famous Face Identification Task: This task contained only a free-fixation component and consisted of 2 blocks, with 100 trials per block.

Human Face Identification Task with 4 Faces: This task contained only a free-fixation component and was run for a single block of 140 trials.

Chimp Face Identification Task: This task contained only a free-fixation component and was run for a single block of 140 trials.

Luggage Bag Identification Task: This task contained only a free-fixation component and was run for a single block of 140 trials.

Experiment Group 2.

Human Gender Discrimination Task: This task consisted of a free-fixation component which was run for 2 blocks of 125 trials each, followed by a forced-fixation component which was run for 10 blocks of 200 trials each (~400 trials for each of 5 forced-fixation positions).

Car Discrimination Task: This task consisted of a free-fixation component which was run for 2 blocks of 125 trials each, followed by a forced-fixation component which was run for 10 blocks of 200 trials each (~400 trials for each of 5 forced-fixation positions).

Experiment Group 3.

Human Face Identification Task with 10 Faces: The forced-fixation component of this task consisted of 1 practice block with 12 trials, followed by 7 experimental blocks with 150 trials each. This was preceded by a free-fixation component (described in Experiment Group 1), which had to always be run before the forced-fixation component in order to obtain an observer-specific preferred fixation position, which was then used as one of the 5 forced-fixation points in the forced-fixation component

Single Gabor Detection Task: This task contained only a forced-fixation component. The first part of the task was a staircase procedure that consisted of 6 blocks of 80 trials each at a single fixation position centered on a gabor. The staircase procedure was done to find observer-specific contrast values for the gabor to set performance at an 85% threshold level. This was followed by the main experimental task, which consisted of 14 blocks, with 105 trials per block (~15 trials for each of 7 forced-fixation positions). This resulted in ~ 210 trials total for each of 7 forced-fixation positions.

Natural Images Matching Task: This task contained only a forced-fixation component

and consisted of 1 practice block with 20 trials, followed by 14 experimental blocks, with 105 trials per block (~15 trials for each of 7 forced-fixation positions). This resulted in ~ 210 trials for each of 7 forced-fixation positions.

Experimental Conditions.

Here we explain aspects of a general trial timeline for free-fixation and forced-fixation tasks because many of the tasks have very similar trial timelines. We then explain what is different about specific tasks that may deviate from the general trial setup.

Starting Fixation.

Free-Fixation.

As shown in the lower left boxes of Figure 1(a-i), during free-fixation tasks, participants started a trial by pressing the space bar while fixating a cross (.44 º x .44 º) in one of eight randomly chosen locations located on average 13.94º from the center of the stimulus. For all free-fixation tasks, the fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants moved their eyes more than 1º from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus.

Forced-Fixation.

As shown in the lower right boxes of Figure 1(a,h,i), during forced-fixation blocks of those tasks, the cross was located in one of 5 locations, which corresponded to the forehead, eyes, nose, and mouth every 5.07 º downward respectively, and a fifth point that corresponded to an individual-specific preferred fixation position that was found during a

free-fixation condition.

For the natural scenes matching task shown in Figure 1j, the forced-fixation cross was located in one of 7 locations that included a location in the center of the screen and 3 locations located every 1.705 º above and below the central location.

For the single gabor detection task shown in Figure 1k, the forced-fixation cross was located in one of 7 locations that included a location in the center of the screen and 3 locations located every 3 º above and below the central location. For all forced-fixation tasks, the fixation cross was displayed for a random period of time between 500ms and 1500ms to prevent anticipatory eye movements. If participants moved their eyes more than 1º from the center of the fixation cross before the stimulus was displayed or while the stimulus was present during the forced fixation condition, the trial would abort and restart with a new stimulus.

Trial Timing.

Free-Fixation.

As shown Figure 1(a-i), for free-fixation tasks, a stimulus was presented after an initial fixation cross and was shown for 350ms, followed by a noise mask for 500ms (Figure 1(a-g)), or 250ms (Figure 1 (h-i)). For each task, on each trial the stimulus identity that was shown was taken from a uniform distribution (equal probability) of all stimulus identities. This was then followed by a response screen presented for an unlimited time until a participant response. Feedback was given in all free-fixation tasks, except the Gender Discrimination and Sports/Regular Car Discrimination tasks in Figure 1 (h-i). Feedback was not given in those two tasks because they were done by the same participants in an interleaved fashion and giving feedback in the Sports/Regular Car Discrimination may have

biased participants to start doing the task based on the identification of specific cars rather than car categories.

Forced-Fixation.

As shown Figure 1(a,h,i), for forced-fixation tasks, a stimulus was presented after an initial fixation cross and was shown for 200ms. For each task, on each trial the stimulus identity that was shown was taken from a uniform distribution (equal probability) of all stimulus identities. The rest of the trial timeline for these tasks was the same as described above for the free-fixation versions of them. The short presentation time for the forced-fixation condition in the tasks referenced above was used in order to account for the fact that participants did not need time to make an eye-movement from the periphery of the screen, as they did in the free-fixationcondition, as well as to make the task more difficult in order to avoid ceiling effects.

For the natural scenes matching task shown in Figure 1j, a 5x5 grid of small squares with different natural scenes was then shown in the center of the screen after a starting forced-fixation. There was a 50% probability on each trial that the scene in the square at the top center of the grid was the same one as the scene in the square at the bottom center of the grid. At the end of a trial, participants selected with the mouse whether the top and bottom squares in the vertical center of the grid were “matched” or “unmatched” and then immediately received feedback.

For the single gabor detection task shown in Figure 1k, a cue was then briefly shown in the center of the screen after a starting forced-fixation to guide the participants’ attention to that location. A single gabor was then shown in the center of the screen with a 50% probability on each trial at the location of the cue. A noise mask followed after that to limit

visual and cognitive aftereffects. At the end of a trial, participants selected with the mouse whether a gabor was “present” or “absent” in the center of the screen during the trial and then immediately received feedback.

Ideal Observer Models.

Foveated Ideal Observer (FIO) Model.

Here we run several different variants of an Ideal Observer model. First, we describe the main model that we use, which is the FIO. We run the FIO model on three different tasks: a human face identification task, a human face gender discrimination task, and a sports/regular car discrimination task. Below, we briefly give an overview of the algorithmic procedures of running the model for a face identification task (see Appendix for details, as well as an explanation of how this model can be generalized to the two other tasks mentioned above).

A spatially variant contrast sensitivity function (SVCSF) was used to model the degradation of the quality of information obtained in the periphery of a foveated visual system (M. F. Peterson & Eckstein, 2012):

$$SVCSF(f, r, \theta) = c_0 f^{a_0} \exp(-b_0 f - d_0(\theta) r^{n_0} f) \qquad (4.1.1)$$

where *f* is spatial frequency in cycles per degree of visual angle. The terms a_0, b_0, and c_0, were chosen constants set to 1.2, 0.3, and 0.625 respectively, to set the maximum contrast at 1 and the peak at 4 cycles per degree of visual angle at fixation. The polar coordinates *r* and θ specify the distance in visual angle and direction from fixation. d_0 specifies the eccentricity factor as a function of direction, which represents how quickly information is degraded in the periphery. n_0 specifies the steep eccentricity roll off factor. In the model simulations, different parameters are used for d_0 for the vertical up, *du*, vertical down, *dd*, and horizontal, *dh*, directions. The parameters *du*, *dd* , *dh*, and n_0 were previously fit to the forced-fixation condition in a face identification task with a group of 25 participants in (M. F. Peterson &

Eckstein, 2012) in order to match human performance (proportion correct) as a function of fixation position (4 different fixations down the vertical midline of the face). The values used for parameters *du*, *dd*, *dh*, and η_0 respectively, are 2E-6, 9E-6, 1E-6, and 5. The Akaike Information Criterion (Akaike, 1974), which takes into account the variance for each data point, is used as a distance measure, which was minimized between human forced-fixation performance at the four points mentioned above and the model's performance at the same four points. The same parameters are used for the gender discrimination task and regular/sports car discrimination task. (see Human Psychophysics Studies above). The circular plots between Figure 2a and Figure 2b show examples of 2d contrast

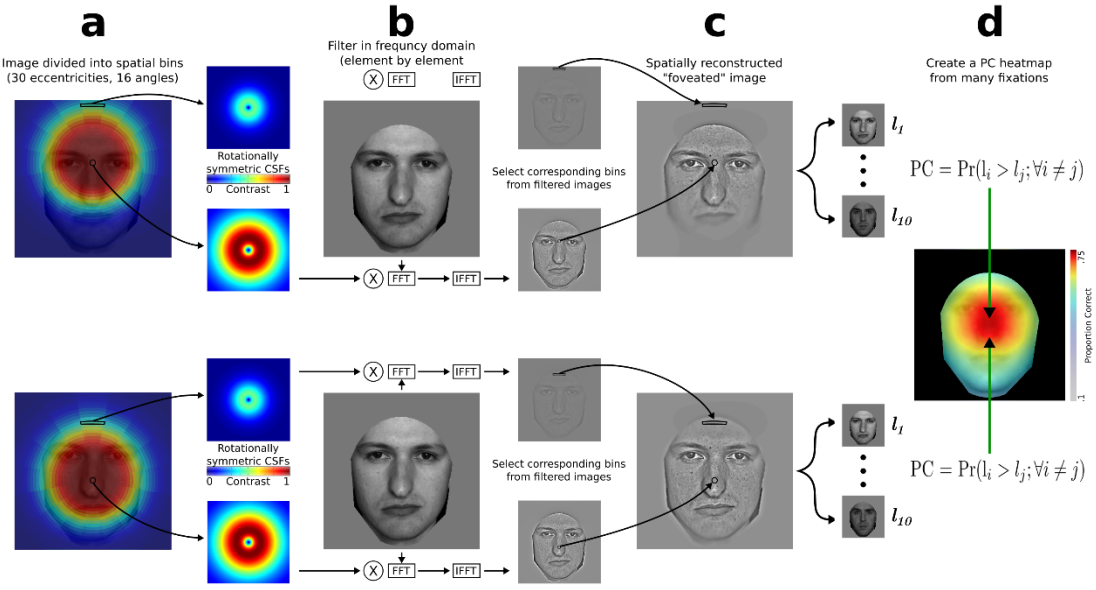

Figure 2: A summary of the process of the computations in the FIO for two fixation positions. The top panels show a fixation point that is below the eyes, which is optimal in a human face identification task. The bottom panels show a fixation that is above the tip of the nose, which is suboptimal for a human face identification task. ***(a)*** *Many trials are simulated where on each trial, a face template is chosen as a signal. The chosen face image on a particular trial is conceptually divided into bins that correspond to specific CSFs as a function of retinal eccentricity. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the CSF plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low-contrast blue in the CSF plots).* ***(b)****,*

The image is transformed into the frequency domain, filtered separately by each possible CSF (here only two are shown), and then transformed back into the spatial domain, resulting in a set of differently filtered images corresponding to each bin. ***(c)****, Corresponding bins are then extracted from the filtered images and input into a composite image that simulates foveation.* ***(d)*** *A set of response variables are then calculated, from which a set of likelihoods is found of each face identity given the noisy image input. A decision of which face was shown is made by taking the maximum likelihood. The FIO model is also run for the face gender discrimination task and sports/regular car discrimination task separately, where likelihoods are summed across exemplars within a class and then the maximum is taken across two summed class likelihoods, rather than across all individual exemplars. Across many trials, a set of proportion correct (PC) values is found, one for each fixation point, and then combined into a heatmap. iFFT, Inverse FFT.*

sensitivity functions at 2 different locations with respect to the fixation position. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low contrast in blue in the plots).

Here, we first describe the algorithmic procedures of running a 1 of 10 face identification task. We simulate many trials of this task. On each trial of the simulation, the face templates $\{\mathbf{f}_1,\ldots,\mathbf{f}_n\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen, where *n* is 10 for the face identification task. Each face template, $\mathbf{f}_i$, consists of a 500x500 pixel face image that is normalized for the position of the eyes and chin as well as for contrast (see Stimuli section of Human Psychophysics Studies above). The same contrast and additive white noise that was used for psychophysics experiments in humans is then added to a chosen template, *i*, before being linearly filtered with the SVCSF and corrupted with additional internal white noise to become the input data, $\mathbf{g}_k$, to the ideal observer:

$$\mathbf{g}_k = \mathbf{E}_k(\mathbf{s}_i + \mathbf{n}_{ex}) + \mathbf{n}_{in} \qquad (4.1.2)$$

where *k* indexes a specific fixation position that serves as the center of a foveation simulation, $\mathbf{n}_{ex}$ is the external Gaussian white noise, $\mathbf{n}_{in}$ is the internal Gaussian white noise, and $\mathbf{E}_k$ is the linear operator that simulates the fixation dependent foveation of the input. This foveated signal is compared (by taking a dot product) to similarly foveated noiseless templates (original face images) to arrive at a set of responses, $\mathbf{r}_{f,k}$, which come from a multivariate Gaussian distribution with a known mean, $\boldsymbol{\mu}_{f,k}$, and covariance matrix, $\sum_k$ (see Appendix for details on how they are calculated):

$$\mathbf{r}_{f,k} \sim MVN(\boldsymbol{\mu}_{f,k}, \sum_k) \tag{4.1.3}$$

Using Bayes rule, the FIO finds a set of posterior probabilities, one for each hypothesis that face f was shown, H_f , given a set of responses $\mathbf{r}_{f,k}$. The posterior probability, $P(H_f \mid \mathbf{r}_{f,k})$, is calculated using the prior probabilities, $P(H_f)$, and the likelihood, $P(\mathbf{r}_{f,k} \mid H_f)$, of the set of responses given the presence of each face, *f*, and the observer's fixation at spatial location, *k*:

$$P(H_f \mid \mathbf{r}_{f,k}) = \frac{P(\mathbf{r}_{f,k} \mid H_f)P(H_f)}{P(\mathbf{r}_{f,k})} \propto P(H_f)P(\mathbf{r}_{f,k} \mid H_f) \tag{4.1.4}$$

The maximum posterior probability is then chosen as the answer:

$$decision = \underset{f}{\operatorname{argmax}}(P(H_f \mid \mathbf{r}_{f,k})) \tag{4.1.5}$$

Bayesian Ideal Observer.

Here, we describe the computations involved for a basic Ideal Observer model, which utilizes image information to achieve the highest possible performance and does not simulate the foveation of the visual system like the FIO described below. Just like for the FIO, we run

a 1 of 10 face identification task with a set of 10 front-view Caucasian male face images that are normalized for the position of the eyes and chin as well as for contrast (see the Stimuli subsection of Human Psychophysics Studies above for details), as well as a face gender discrimination task and a sports/regular car discrimination task. Here we describe the algorithmic details for the face identification task, but the algorithm can be generalized to a 2-class problem using marginalization just like with the FIO (see *Generalization to a 2-Class Problem Using Marginalization* section of the Foveated Ideal Observer Model description in the Appendix).

On each trial of the simulation, the face images $\{\mathbf{f}_1,\ldots,\mathbf{f}_{10}\}$ are sampled uniformly at random and a face template, $\mathbf{s}_i$, is chosen. The same contrast and additive white noise that was used for humans is then added to a chosen template, *i*. The input data, $\mathbf{g}$, to the ideal observer on each simulated trial is then the sum of a random (1 of 10) face template, $\mathbf{s}_i$, and external noise, $\mathbf{n}_{ex}$.

$$\mathbf{g} = \mathbf{s}_i + \mathbf{n}_{ex} \tag{4.2.1}$$

The ideal observer does not have any sources of suboptimality such as internal noise or filtering operations on the face template, $\mathbf{s}_i$, that models foveation. Using Bayes rule, the ideal observer finds a set of posterior probabilities, one for each hypothesis, H_f, that face, *f,* was shown, given the image data, $\mathbf{g}$. The posterior probability, $P(H_f \mid \mathbf{g})$, is calculated using the prior probabilities, $P(H_f)$, and the likelihood, $P(\mathbf{g} \mid H_f)$, of the image data, $\mathbf{g}$, given the presence of each face, *f*:

$$P(H_f \mid \mathbf{g}) = \frac{P(\mathbf{g} \mid H_f)P(H_f)}{P(\mathbf{g})} \propto P(H_f)P(\mathbf{g} \mid H_f) = l_f \tag{4.2.2}$$

The normalizing factor, $P(\mathbf{g})$, in equation (4.2.2) is the same for all posterior probabilities, so it can be ignored without changing the result. The likelihood, $P(\mathbf{g} \mid H_f)$, of the signal having come from a particular face is calculated from a known distribution that comes from a product of distributions of individual pixel noise (see Appendix for details).

The maximum posterior probability is then chosen as the answer:

$$decision = \underset{f}{\operatorname{argmax}}(P(H_f \mid \mathbf{g})) \tag{4.2.3}$$

Convolutional Neural Network (CNN) Model.

Stimuli.

Training.

The CNN model is separately trained on two different face tasks: a human face identification task and a human face gender discrimination task. The training stimulus set for the face identification task is created from a base set of 10 face images that are used in the face identification task of the human psychophysics experiments (see Stimuli section of Human Psychophysics Experiments above). A base training set of 4000 face images was created by making 400 copies with added Gaussian white noise of each of the 10 original face images. Multiple copies of the 4000 training face images were then made by adding a simulation of the foveation of the visual system centered at different fixation positions as shown in Figure 3a below. The fixation positions correspond to the forehead, eyes, nose, and mouth. These are the same fixation positions that are used in the human forced-fixation condition of the face identification task. To create the training stimulus set for the gender discrimination task, a set of ~3500 face photos were first scraped from the Humane project website (http://humanae.tumblr.com/) of photographer Angelica Dass. This stimulus set has been previously used to train generative models of photorealistic faces (Suchow, Peterson, & Griffiths, 2018) because the faces are all front-facing, with controlled lighting, and come from a variety of ages and ethnicities. After scraping the faces from the Humane project website, they were then spatially aligned such that the face was positioned with center of the eyes at 2/5 of the image height below the top of the image and with the chin 1/50 of the image height above the bottom of the image. This was done by extracting facial landmarks around the eyes and mouth using the Python dlib library and then rotating, resizing, and

cropping the images to a size of 224x224 pixels. The face images were then manually labeled for gender as well as filtered for age, facial hair, and excessive head hair. Faces that looked like they were younger than age 18 were discarded as well as faces of either gender that had one or more major face features covered with head hair. Male faces with more than a mild amount of facial hair were also discarded in order to not bias the CNN model during training, to avoid facial hair becoming an important feature in the gender discrimination task. This resulted in a disproportionate number of male face images being discarded because of excessive facial hair, leaving 1541 female faces and 616 male faces. 80% of the female and male faces were then randomly selected for a training set. However, due to a disproportionate number of female faces remaining relative to the number of male faces, the male faces were randomly oversampled such that two copies were made of each male face and then random subset of the original male faces were chosen for a third copy to create a total number of male images equal to the total number of female images for training. This resulted in a total of 1233 male and 1233 female face images. Four different sets of the training images were then created by foveating them in the same was as the face images in the face identification task.

Testing.

The testing stimuli sets for the face identification task were created in the same way as the training stimuli sets, except a total of 1000 face images were used, with 100 noisy copies for each of 10 face identities. Four sets of the 1000 face images were also created by simulating foveation at the same locations as described for the training sets above. The testing sets for the face gender discrimination task were created from the remaining 20% of the face images that were not used for the training set, resulting in 308 female and 308 male

images, after doing the same oversampling procedure for the male images, as described for the training set above. Four sets of the 1000 face images were also created by simulating foveation at the same locations as

Architecture and Settings.

Original resnet-18.

We use an 18-layer resnet-18 (K. He et al., 2015) architecture to separately run both a 10-class human face identification task and a 2-class human face gender discrimination task. The network is made up of 4 "residual blocks," each of which contain 2 pairs (this number is higher for other variants of this network structure) of the same layer structure (same size and depth of feature maps) (Figure 3b). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem.

We use mini-batch (200 images per batch) stochastic gradient descent (SGD) along with a cross-entropy loss function to optimize the parameters in the model. We use hyperparameter settings of 5e-4 for the learning rate and .9 for momentum. Although this network can theoretically be run with any input image size, here we run it with an image size of 224x224 pixels which is a standard image size that is used to train common CNN architectures.

Modified resnet-18 and Class-Specific Activations Visualization.

In addition to the original resnet-18 network, we use also the methodology of (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) and run a modified version of the same network in order to be able to construct a visualization of the important features in the input stimuli

that are used by the network to do the emotion classification task. This is done by mapping a weighted linear combination of the 14x14 feature maps of the last convolutional layer of the network onto the original 224x224 input images. The weights used to combine the feature maps come from the learned connections between the Global Average Pooling (GAP) layer, which acts as a unidimensional representation of the 14x14 feature maps preceding it, and the class scores output by the network. For each of the three classes, a specific Class Activation Map (CAM) is found by using the weights connecting the GAP layer to a specific class. Although (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015) used this method for localization of objects in complex classification tasks with a large number of classes, it is still useful for our purpose of visualizing the features of faces that are most discriminative for the network during this task. Since the faces are aligned during both the training and testing phases, the discriminative features should be located in specific areas across CAMs. For the same reason, the CAM's can be averaged across classes to get a single CAM for a specific combination of training and testing sets and specific task.

Figure 3c shows the modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering the stride from 2 to 1 during the convolution operation before the last residual block. The difference in the modified network is outlined in red.

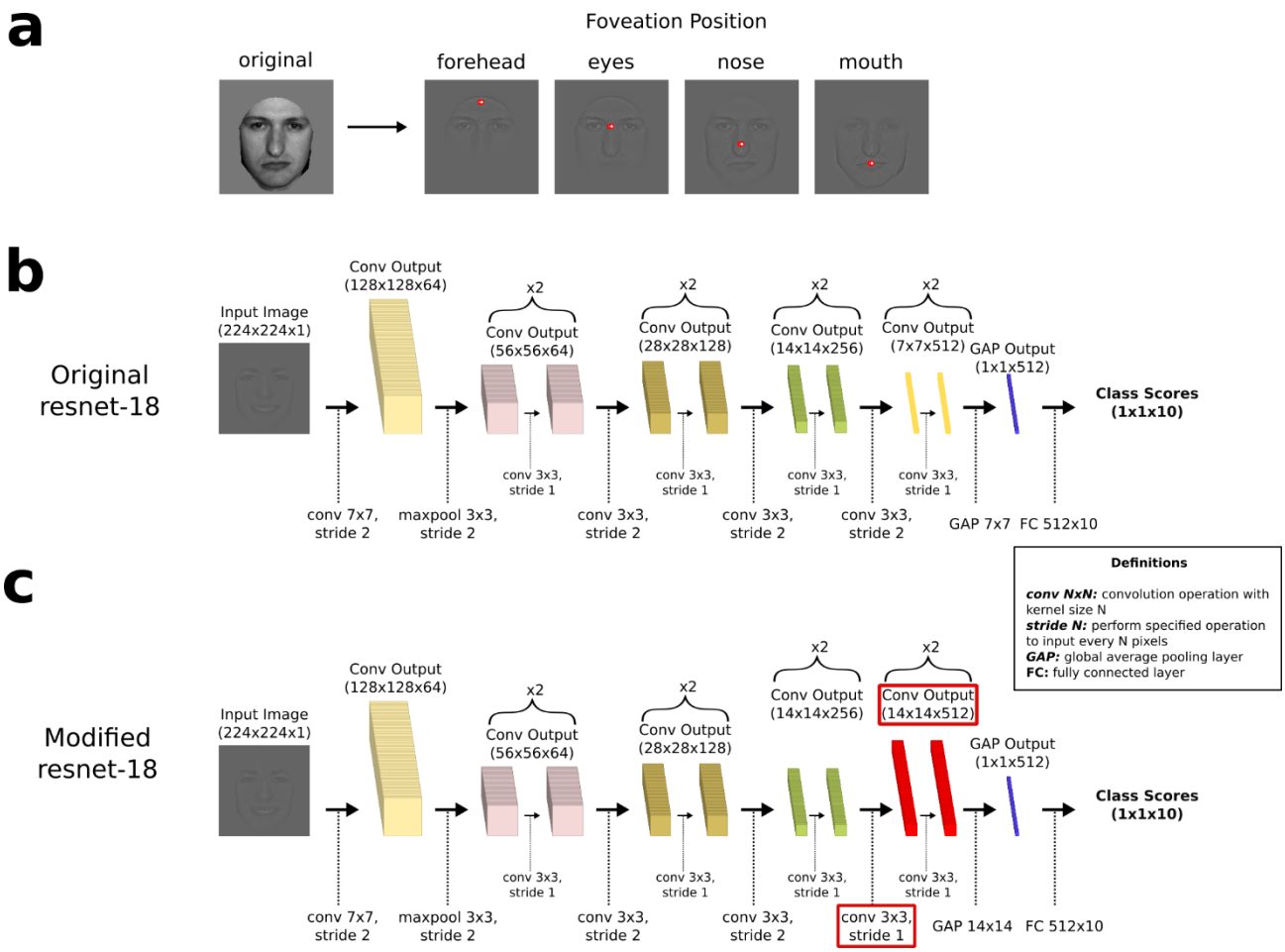

*Figure 3: (**a**) An example of a face image being processed with spatial filtering operations that simulate foveation at four different positions along the vertical midline of a face. Four "foveated" versions of each of the training and testing sets were created to simulate possible differences in human development of an internal face template representation. (**b**) The top flowchart shows the structure and operations involved in the original resnet-18 network* (K. He et al., 2015). *Although this network can theoretically be run with any input image size, here we run it with an image size of 224x224 pixels and show the sizes of feature map outputs after max pooling and convolution operations along with the chosen depths of the feature maps at each layer, which are fixed parameter settings. Similarly, although the network is able to learn to classify an arbitrary number of classes, here we show an output of class scores for a 10-class face identification task. One aspect of the resnet network that isn't explicitly shown in the flowchart is the "skip-connections" between layers of the same size. The network is made up of 4 "residual blocks," each of which contain 2 pairs of the same layer structure (same size and depth of feature maps). In cases where it is more advantageous to do so, the network is able to learn an identity mapping between consecutive layers of the same size within a residual block, which in essence allows the network to skip layers if needed, and tune itself to a network size that is optimal for a specific classification problem. (**c**) The bottom flowhart shows a modified version of the resnet-18 network, where the feature maps (height and width, but not depth) of the fourth residual block are larger. Implementing the change relative to the original resnet-18 network only involves lowering*

the stride from 2 to 1 during the convolution operation before the last residual block. The difference in the modified network is outlined in red. We implement this modification in order to output a set of 14x14 pixel feature maps instead of 7x7 pixel feature maps. This allows us to use the methodology of (Zhou et al., 2015) *to construct a visualization of the important features in the input stimuli that are used by the network to do the emotion classification task. This is done by mapping a linear combination of the 14x14 feature maps onto the original 224x224 input images.*

4.3 Results

Individual Differences in Initial Fixation Position Exist Across Human Population.

First, in order to identify individual differences in initial fixation position to faces between participants, we show a histogram with locations and frequencies of average vertical initial fixation positions to a face during a face identification task from a database of 285 observers in Figure 4a. The database of observers was compiled from initial fixation location data presented in this paper as well as data previously presented in (Or et al., 2015; M. F. Peterson & Eckstein, 2012, 2014; M. F. Peterson et al., 2016). The mode of the vertical fixation distribution is at a point that is a bit below the eyes, and there is a long tail going down toward the mouth region. For most observers, the variability of the initial fixation position across trials is fairly small. In Figure 4b, an example of fixation positions for three different observers that we classify as eye-lookers, nose-lookers, and mouth-lookers, respectively, is shown with fixation positions from individual trials overlaid onto an average of the face stimuli used for a face identification task. Although the vertical fixation distribution in Figure 4a is unimodal, we choose a boundary of separation of observers into eye-lookers and nose-lookers based on the distance of their average vertical fixation position from the location of the eyes in the spatially aligned face stimuli. We set that distance to be 50% of the distance from the eyes to the tip of the nose, which corresponds to 1.776 degrees below the eyes, in a face stimulus that is 17.8 degrees in height (not including the portion of the stimulus that is covered by a mask in order to hide the hairline, ears, and neck). Due to the small number of mouth-lookers, we group them under nose-lookers and focus only on

differences between two groups: eye-lookers and everyone else, who we group under nose-lookers. This allows us to then compare performance profiles in both human-face and non-human-face tasks between the two groups of participants in order to determine what may be the cause of the observed individual differences.

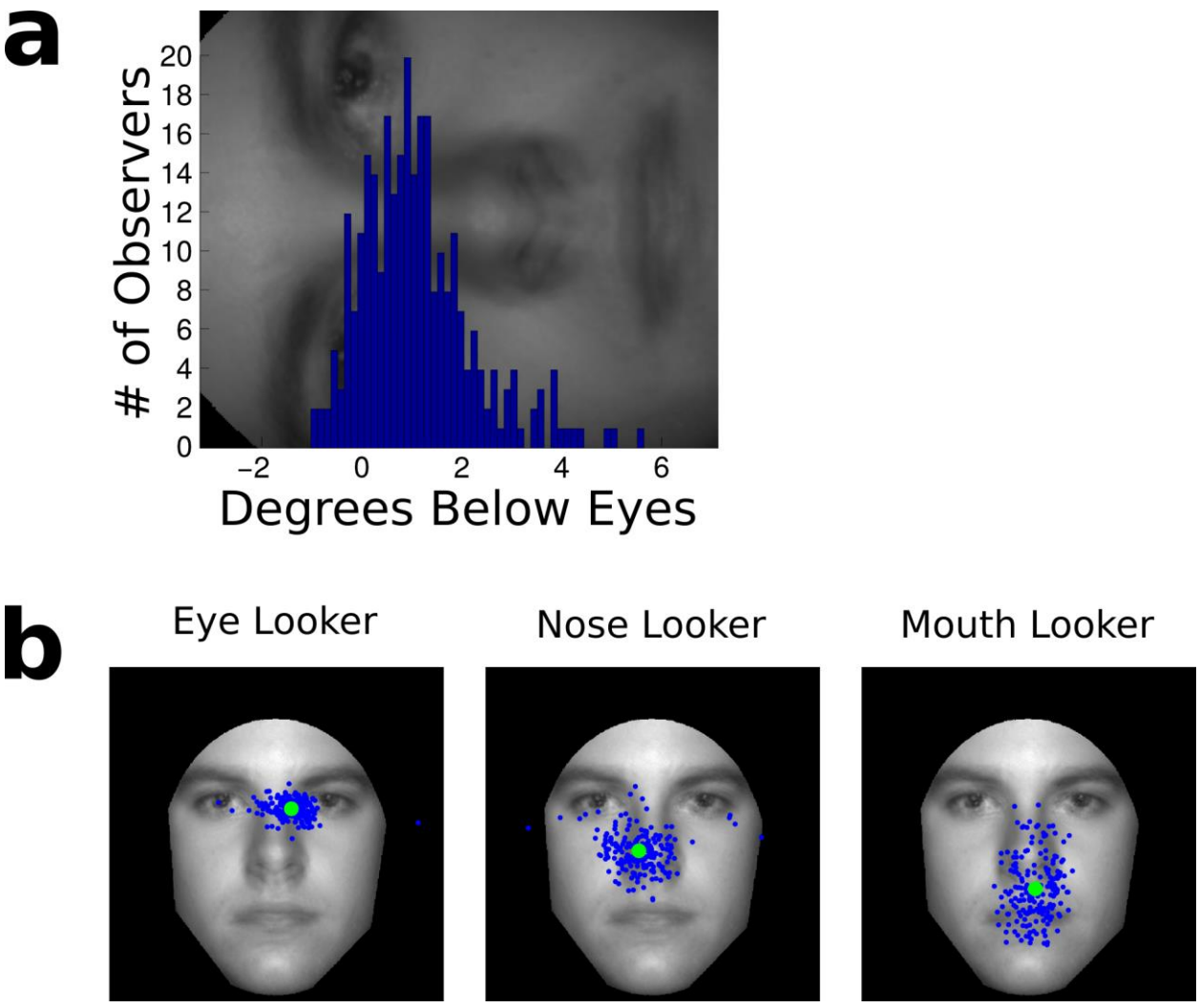

*Figure 4: **a**) Here a histogram is shown with locations and frequencies of average vertical initial fixation positions to a face during a face identification task from a database of 285 observers. The mode of the vertical fixation distribution is at a point that is a bit below the eyes, and there is a long tail going down toward the mouth region. **b**) An example of fixation positions for three different observers that we classify as eye-lookers, nose-lookers, and mouth-lookers, respectively, is shown with fixation positions overlaid onto an average of the face stimuli used for a face identification task. The blue points represent fixation positions from individual trials and the green points represent averages of the blue points. Although the vertical fixation distribution in (a) is unimodal, we choose a boundary of separation of*

observers into eye-lookers and nose-lookers based on the distance of their average vertical fixation position from the location of the eyes in the spatially aligned face stimuli. Due to the small number of mouth-lookers, we group them under nose-lookers and focus only on differences between two groups: eye-lookers and everyone else, who we group under nose-lookers.

Comparison of Free-Fixation Metrics across Human-Face and non-Human-Face Tasks Show Higher Consistency in Fixation Strategies in Human-Face Tasks.

In order to determine human initial fixation consistency across human face discrimination tasks, we ran participants in a free-fixation condition in a human face identification task, a human face gender identification task, a famous faces identification task, and an emotion identification task. Figure 5a shows that there is a high correlation in the initial vertical fixation position between the human face identification task compared to each of the tasks mentioned above. The correlation values for the vertical fixation position between the human identification task compared to the gender identification, famous faces identification, and emotion identification task, respectively are 0.84, 0.72, and 0.85. All correlation values are significant and show a high consistency in fixation behavior between different face identification tasks ($t(23) = 7.51$, $p = 1.24E-7$; $t(23)= 4.96$, $p = 5.17E-5$; $t(23) = 7.74$, $p = 7.56E-8$) for the human identification task compared to the gender identification, famous faces identification, and emotion identification task, respectively. In addition to participants completing a free-fixation task with human faces, a different set of participants completed a free-fixation condition in a human face identification task, a chimp face identification task, and a luggage bag identification task. Figure 5b shows correlation values for the vertical fixation position in the human identification task compared to the chimp face identification task and luggage bag identification task, which are 0.71 and .50, respectively. Both correlation values listed above are significant ($t(76) = 8.79$, $p = 3.04E-11$; $t(76) = 5.06$, $p = 7.96E-6$), but the correlation for luggage bag identification is significantly lower than for chimp face identification ($z(78) = 2.07$, $p = 0.0385$, two-tailed, using the Fisher z-

transformation). In addition to comparing vertical fixation positions between a human face identification task, a chimp face identification task, and a luggage bag identification task, we also measure differences in the variance of the vertical fixation positions across tasks. Figure 5c shows box and whisker plots of variance distributions of the vertical initial fixation position in the human face identification, chimp face identification, and luggage bag identification task are shown, respectively. The variances are significantly different across the three tasks, with higher variance in the chimp face identification task compared to the human face identification task ($p < 1E-4$), and higher variance in the luggage bag identification task compared to both the human face ($p < 1E-4$) and chimp face identification tasks ($p<1E-4$). The statistical test used to find the differences in variances was done with a bootstrap procedure by creating 10,000 samples from each of the three original variance samples (humans, chimps, and bags) and making pairwise comparisons of empirical distributions of variance means using 10,000 percentile values. Figure 5d also shows example fixation positions for the same single observer performing a human face identification task, a chimp face identification task, and a luggage bag identification task, respectively.

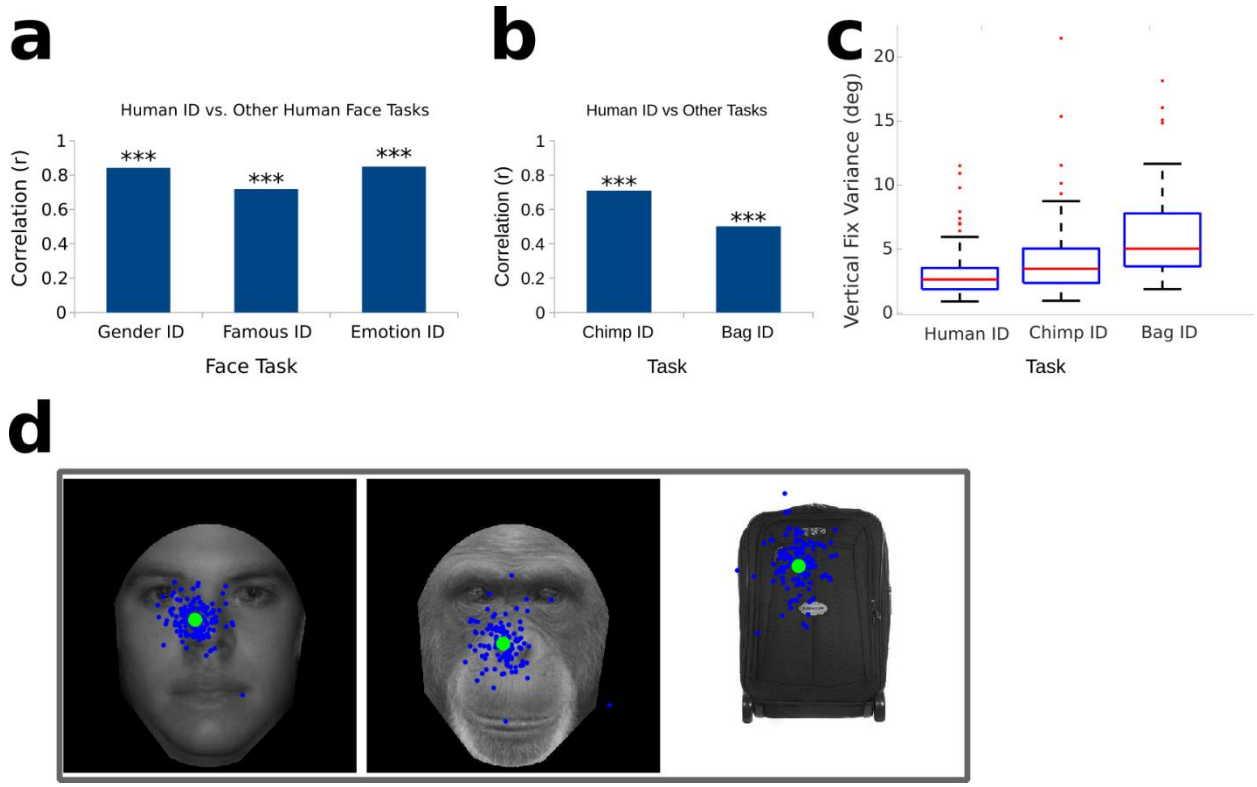

Figure 5: ***a****) Correlation values are shown for the vertical fixation position between the human identification task compared to the gender identification, famous faces identification, and emotion identification task, respectively. All correlation values are significant and show a high consistency in fixation behavior between different face identification tasks.* ***b****) Correlation values are shown for the vertical fixation position between the human identification task compared to the chimp face identification task and luggage bag identification task. Correlations values are significant for both tasks being compared to human face identification, but the correlation for luggage bag identification is significantly lower than for chimp face identification.* ***c****) Box and whisker plots of variance distributions of the vertical initial fixation position in the human face identification, chimp face identification, and luggage bag identification task are shown, respectively. The red horizontal lines in the center of each plot denote the mean variances for each task, and the blue horizontal lines below and above the red lines denote the 25th and 75th percentiles of the variance distributions for each task. The differences in the variances are significantly different across the three tasks, with higher variance in the chimp face identification task compared to the human face identification task, and higher variance in the luggage bag identification task compared to both the human face and chimp face identification tasks. The statistical test used to find the differences in variances was done with a bootstrap procedure.* ***d****) Example fixation positions are shown for the same observer performing a human face identification task, a chimp face identification task, and a luggage bag identification task, respectively. The blue points are fixations from a single trial and the green points are averages of the blue points.*

Lack of Generalizability of Fixation Strategies and Lack of Efficient Information Use in non-Face Tasks.

Here we compare human preferred points of initial fixation during a free-fixation condition in a human gender discrimination task vs. a sports/regular car discrimination task. In addition, we measure performance profiles of participants in a forced-fixation condition for both tasks to determine the existence and location of empirical optimal points of fixation. We then compare the human performance data to performance profiles found with an FIO model, which also determines the location of theoretically optimal points of initial fixation for both tasks. Figure 6a shows an FIO performance map, which contains proportion correct values for every possible fixation position to an image, for a gender identification task. The part of the performance map that is the darkest red, represents the highest performance, which is a theoretically optimal point of initial fixation for the gender identification task. The average human preferred point of initial fixation, shown in green, is very close to the theoretical optimal point. In Figure 6b, the performance of the FIO model down the vertical midline of the face image is shown for a gender identification task. Human performance data from the forced-fixation condition of the same task at different points down the vertical midline of the face image is also shown. The model's parameters were previously fit to a human face identification task with a different set of observers. Here only the internal noise parameter was fit to the human forced-fixation performance data. The human preferred point of fixation is close to the human empirically optimal point of fixation, which is predicted well by the theoretically optimal point found with the FIO model. The overall human forced-fixation performance profile is also predicted relatively well by the model, with an Akaike Information Criterion (AIC) score of 2.83.

In contrast to the ability of the FIO model to generalize between face tasks, having been originally fit to human data in a face identification task, and predicting the human performance profile well in a gender discrimination task, the FIO does not generalize well to non-face tasks. The same parameters of the FIO model were used in the sports/regular car identification task as in the gender identification task, except the internal noise value, which was fit to the human forced-fixation performance data. As shown in Figure 6c and Figure 6d, the although the FIO performance profile is able to provide a relatively close fit to human data for a sports/regular car discrimination task, with an AIC score of 2.09, the average human preferred fixation position is far from the location of the theoretical optimal point.

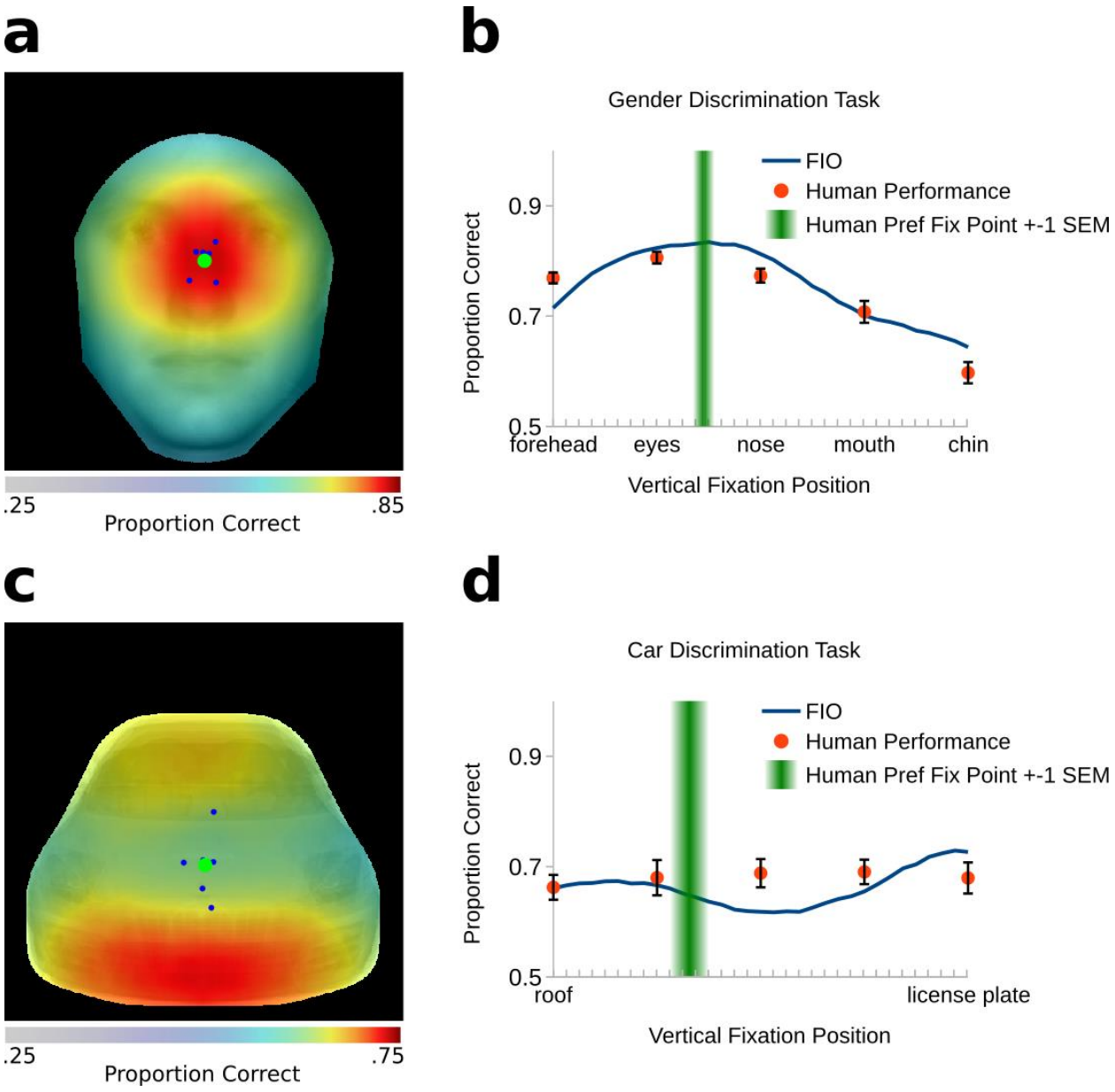

Figure 6: ***a****) An FIO performance map, which contains proportion correct values for every possible fixation position to an image, is shown for a gender identification task and is overlaid on top of a face image. Average fixation positions from six observers for the same task are overlaid as blue points, along with an average across observers in green. The part of the performance map that is the darkest red, represents the highest performance, which is a theoretically optimal point of initial fixation for the gender identification task. The average human preferred point of initial fixation is very close to the theoretical optimal point.* ***b****) The performance of the FIO model down the vertical midline of the face image is shown in blue for a gender identification task. Human performance data from the forced-fixation gender identification task at different points down the vertical midline of the face image is shown in orange, with error bars representing one standard error above and below the mean. The green bar shows the location of the preferred vertical fixation position to a face averaged*

across observers (same point shown in green in (a)) for this task, with the width representing one standard error above and below the mean. The model's parameters were previously fit to a human face identification task with a different set of observers. Here only the internal noise parameter was fit to the human forced-fixation performance data. The human preferred point of fixation is close to the human empirically optimal point of fixation, which is predicted well by the theoretically optimal point found with the FIO model. The overall human forced-fixation performance profile is also predicted well by the model. ***c****) Similar to (a), a performance map of the FIO model is shown, except for a regular/sports car identification task, along with preferred fixation positions from the same set of observers who did the gender identification task.* ***d****) Similar to (b), the performance of the FIO model as well as performance of human observers down the vertical midline of the car image in the forced-fixation car identification task is shown. A green bar representing one standard error above and below the average preferred vertical fixation position in this task is also shown. The same parameters of the FIO model were used in the sports/regular car identification task as in the gender identification task, except the internal noise value, which was fit to the human forced-fixation performance data. Unlike in the gender identification task, the FIO does not predict the human forced-fixation performance profile well and the average human preferred fixation position is far from the location of the theoretical optimal point.*

Differences in Performance and Preferred Point of Fixation Between Eye-lookers and Nose-Lookers Cannot Be Explained By Differences in Visual Field Anisotropy.

In order to test the altered-anisotropy theory, we compare forced-fixation performance profiles in a human face identification task between eye-lookers and nose-lookers, to forced-fixation performance profiles of the same participants in a natural scene matching task and a single gabor detection task. The latter two tasks are run in order to determine whether there are any significant differences in vertical anisotropy between eye-lookers and nose-lookers in simpler non-face tasks. In Figure 7a, we show a sketch of the altered vertical anisotropy hypothesis, which illustrates differences in contrast sensitivity between eye-lookers and nose-lookers relative to the center of fixation. Here a higher contrast sensitivity is shown for eye-lookers in the lower visual field and a lower contrast sensitivity in the upper visual field, relative to nose-lookers. Under this hypothesis, if nose lookers fixate the eyes, they get lower quality information from the lower part of the face and higher quality information from the upper part of the face relative to when eye-lookers fixate the eyes. If nose-lookers fixate the nose, they are able to get a similar quality of information from both the upper and lower parts of the face relative to when eye-lookers fixate the eyes. If this hypothesis is true, we also expect to see significant differences in performance between eye-lookers and nose-lookers in non-face tasks when comparing use of the upper visual field vs. the lower visual field to extract task-relevant information. Figure 7b shows a forced-fixation performance profile down the vertical midline of a face stimulus for separate groups of eye-lookers in blue, and nose-lookers in red, in a human face identification task. For each group, performance at 5 points is shown, with 4 of them corresponding to the

forehead, eyes, nose, and mouth. The fifth point corresponds to the average vertical fixation position, within each group, taken from individual preferred points of fixation in the free-fixation condition of each task. Eye-lookers perform better, but not statistically significantly so, at the eyes vs. the nose ($t(12) = 1.73$, $p = 0.11$, one-tailed), while nose-lookers perform significantly better at the nose vs. the eyes ($t(11) = 5.64$, $p = 1.5E\text{-}4$, one-tailed). This result is similar to what was reported in (M. F. Peterson & Eckstein, 2013), where participants in a face identification task were first classified as eye-lookers and nose-lookers. Despite different forced-fixation performance profiles in eye-lookers and nose-lookers in a face identification task, there are no significant performance differences between the two groups when comparing average performance using the upper vs. lower visual fields in a natural scene matching task and a single gabor detection task, as shown in Figure 7c and Figure 7d, respectively. First, as found in previous studies of vertical anisotropy, we find that humans perform better when using their lower visual field to process task-relevant stimuli in both the natural scene-matching task ($t(20)=6.34$, $p = 1.4E\text{-}3$, for performance averaged across fixation points in the upper visual field vs. performance averaged across fixation points in the upper visual field for all participants) and the single gabor detection task ($t(24) = 2.195$, $p = 1.67E\text{-}9$, for performance averaged across fixation points in the upper visual field vs. performance averaged across fixation points in the upper visual field for all participants). However, in both the natural scene matching task and single gabor detection task, we find no significant differences between eye-lookers and nose-lookers when comparing a difference in the average performance at the upper visual field fixations vs lower visual field fixations ($t(19) = 0.724$, $p = 0.478$, for the natural scene matching task; $t(23) = 1.205$, $p = 0.24$, for the single gabor detection task).

In Figures 7e-g, we also show correlation plots for the preferred free-fixation position in the human face tasks vs. a magnitude of performance differences between the upper and lower visual field in the forced-fixation condition of the human face identification task, the natural image matching task, and the single gabor detection task, respectively. The preferred free-fixation position for each participant is taken from an average of the preferred vertical fixation positions in the human face identification task, famous face identification task, and gender identification task. The correlation is high and significant ($t(23) = 5.05$, $p = 4.08E\text{-}5$) only when comparing the the average preferred vertical fixation position to the average vertical visual field performance differences in the forced-fixation condition of the human face identification task, but is insignificant when comparing it to the natural scene matching task ($t(19) = 0.702$, $p = 0.49$) and the single gabor detection task ($t(23) = 1.03$, $p = 0.315$).

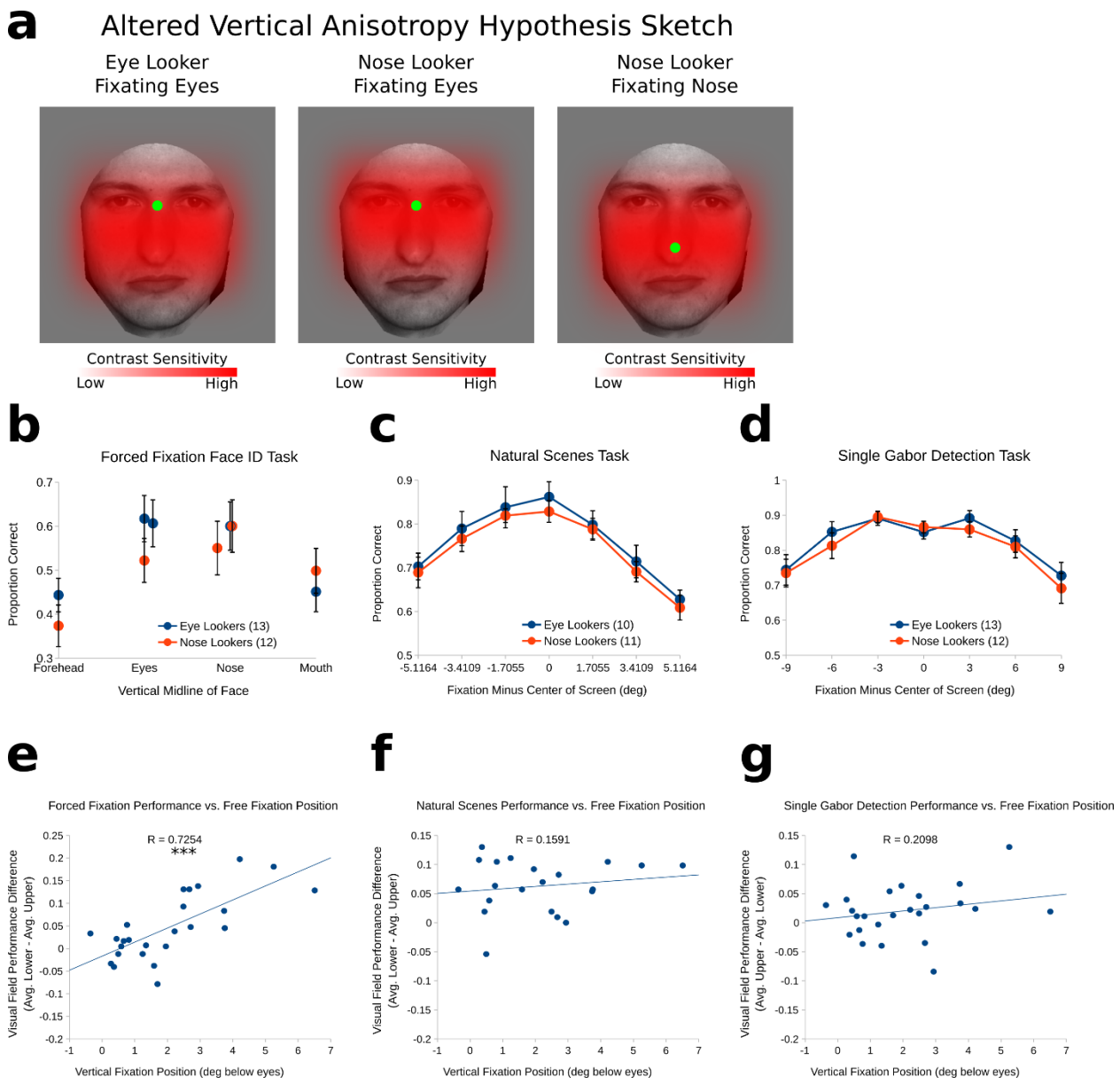

Figure 7: ***a****) A sketch is shown of the altered vertical anisotropy hypothesis, which illustrates differences in contrast sensitivity between eye-lookers and nose-lookers relative to the center of fixation. Here a higher contrast sensitivity is shown for eye-lookers in the lower visual field and a lower contrast sensitivity in the upper visual field, relative to nose-lookers. Under this hypothesis, if nose lookers fixate the eyes, they get lower quality information from the lower part of the face and higher quality information from the upper part of the face relative to when eye-lookers fixate the eyes. If nose-lookers fixate the nose, they are able to get a similar quality of information from both the upper and lower parts of the face relative to when eye-lookers fixate the eyes.* ***b****) A forced-fixation performance profile down the vertical midline of a face stimulus is shown for separate groups of eye-lookers in blue, and nose-lookers in red, in a human face identification task. For each group, performance at 5 points is shown, with 4 of them corresponding to the forehead, eyes, nose, and mouth. The fifth*

point corresponds to the average vertical fixation position, within each group, taken from individual preferred points of fixation in the free-fixation condition of each task. Although an average position is shown for the preferred point, one close to the eyes for the eye-lookers and one close to the nose for the nose-lookers, the actual point used in the forced-fixation task was observer specific. Eye-lookers perform higher, but not statistically significantly at the eyes vs. the nose, while nose-lookers perform significantly better at the nose vs. the eyes. ***c****) Forced-fixation performance in the natural scenes matching task is shown at different fixation positions down the vertical midline of the screen relative to the vertical center. Performance is better on average in both groups when the lower visual field is used more to process task-relevant information, which corresponds to fixation positions in the upper visual field and negative values on x-axis of the plot. However, the ratio of performance when using the upper visual field relative to the lower visual field is not significantly different between eye-lookers and nose-lookers. This task contains less participants because several were removed due to the performing at chance level.* ***d****) Similar to the natural scenes matching task, results of the single gabor detection task are shown for eye-lookers and nose-lookers. Vertical fixation positions that were above the center presentation of the gabor stimulus correspond to negative values on the x-axis of the plot, and use of the lower visual field to process the stimulus. The ratio of performance when using the upper visual field relative to the lower visual field is not significantly different between eye-lookers and nose-lookers in this task.* ***e-g****) Correlation plots are shown for the preferred free-fixation position in the human face tasks vs. a magnitude of performance differences between the upper and lower visual field in the forced-fixation condition of the human face identification task, the natural image matching task, and the single gabor detection task, respectively. The preferred free-fixation position for each participant was taken from an average of the preferred vertical fixation positions in the human face identification task, famous face identification task, and gender identification task. The correlation is high and significant only between the average preferred vertical fixation position and the magnitude of performance difference between the upper and lower visual field in the forced-fixation condition of the human face identification task. The error bars represent one standard error above and below the mean.*

Differences in Performance and Preferred Point of Fixation Between Eye-lookers and Nose-Lookers Can Be Explained By Computational Models of Differences in Internal Face Template Representations.

In order to test the altered face template theory, we run two computational models that are able to represent differences in an internal face template between eye-lookers and nose-lookers and compare them human performance profile data from the forced-fixation condition of the face identification task. The first model that we use is a modified FIO model (FT-FIO; see Methods section for details), where the internal representation of a face is fixed at a specific simulated fixation position. Figure 8a shows the performance profile down the midline of the face for forced-fixation positions at the forehead, eyes, nose, and mouth is shown for eye-lookers and nose-lookers. The performance profile of an FT-FIO model is also shown when the fixed template is at a point below the eyes (blue line) and a separate performance profile for when the fixed template is at the nose (green line). Here, the location of the fixed template represents the internal representation of the foveated face stimuli that the model uses to compare an incoming face stimulus to on each simulated trial. In the FT-FIO model, although the internal representation of the faces is fixed at a specific simulated fixation position, different fixation positions for incoming face stimuli are still tested, which allows us to find a performance profile at different fixation positions. Here, the same parameters are used for the original FIO model, except the internal noise parameters, which is adjusted to best fit the human data for eye-lookers and nose-lookers individually. The FIO model with fixed templates provides a reasonably strong fit (AIC of 1.37 for eye-lookers, and AIC of 0.95 for nose-lookers) to the human data and shows the expected differences in the theoretically optimal points of fixation, which match the location of human preferred points

of fixation and empirically optimal points of fixation. Figure 8b shows the same plot is shown as in Figure 8a, except now all of the parameters of the FIO with a fixed template position below the eyes are fit to the eye-looker human data to provide a slightly better fit (AIC of 0.17 for eye-lookers, and AIC of 0.46 for nose-lookers). The same parameters, except for the internal noise, are then used to run the FIO with a fixed template position at the nose. Similar to the FIO model with the original parameters, the FIO with new parameters also shows the expected differences in the theoretically optimal points of fixation, which match the location of human preferred points of fixation and empirically optimal points of fixation.

The second model that we use to represent differences in an internal face template representation is a CNN model. This model supplements the results of the FT-FIO because we are also able to simulate possible differences in human development that may have led to the different internal face representations. In Figure 8c, the results of a CNN model are shown for a human face identification task, using the same face stimuli as the ones used to run the FIO model and to run the human psychophysics experiments. We simulate differences in human development by using different training sets where the simulated center of foveation is at the forehead, eyes, nose, or mouth. We then test each of the differently trained CNN models with four separate stimulus sets that are again foveated at the four different fixation positions mentioned above. For all four models, performance is highest at the test fixation position that matches the training fixation position. When there is a mismatch in training and testing fixation positions, there is a steep drop in performance, proportional to how far away the training fixation position is from the testing fixation position. Figure 8d shows similar results as in Figure 8c, except these results are now for a

CNN model trained on a face gender discrimination task instead of a face identification task. We run this model in order to show that the matched-template principle of performance being maximized when the training fixation position matches the testing fixation position, generalizes to a different face task with different face stimuli. Since gender discrimination is a two-class problem, the performance profiles for each of the trained models are much less steep than for a face identification task, and chance performance is at a proportion correct value of 0.5 instead of at 0.1. In Figure 8e, visualizations are shown of the parts of the face stimuli that the CNN model trained on a face identification task uses the most, for training fixation positions at the eyes and nose, and testing fixation positions at the forehead, eyes, nose, and mouth. We use the method of (Zhou et al., 2015) to find class activations maps (CAMs), and then average across them to create each of the images seen in Figure 8e. The visualizations show that the eyes are an important region that the CNN uses to do the classification task, regardless of which fixation position was used in the training stimuli. However, when the test fixation location is very far from the training fixation location, the model makes use of other features besides the eyes, and this varies depending on the training fixation location.

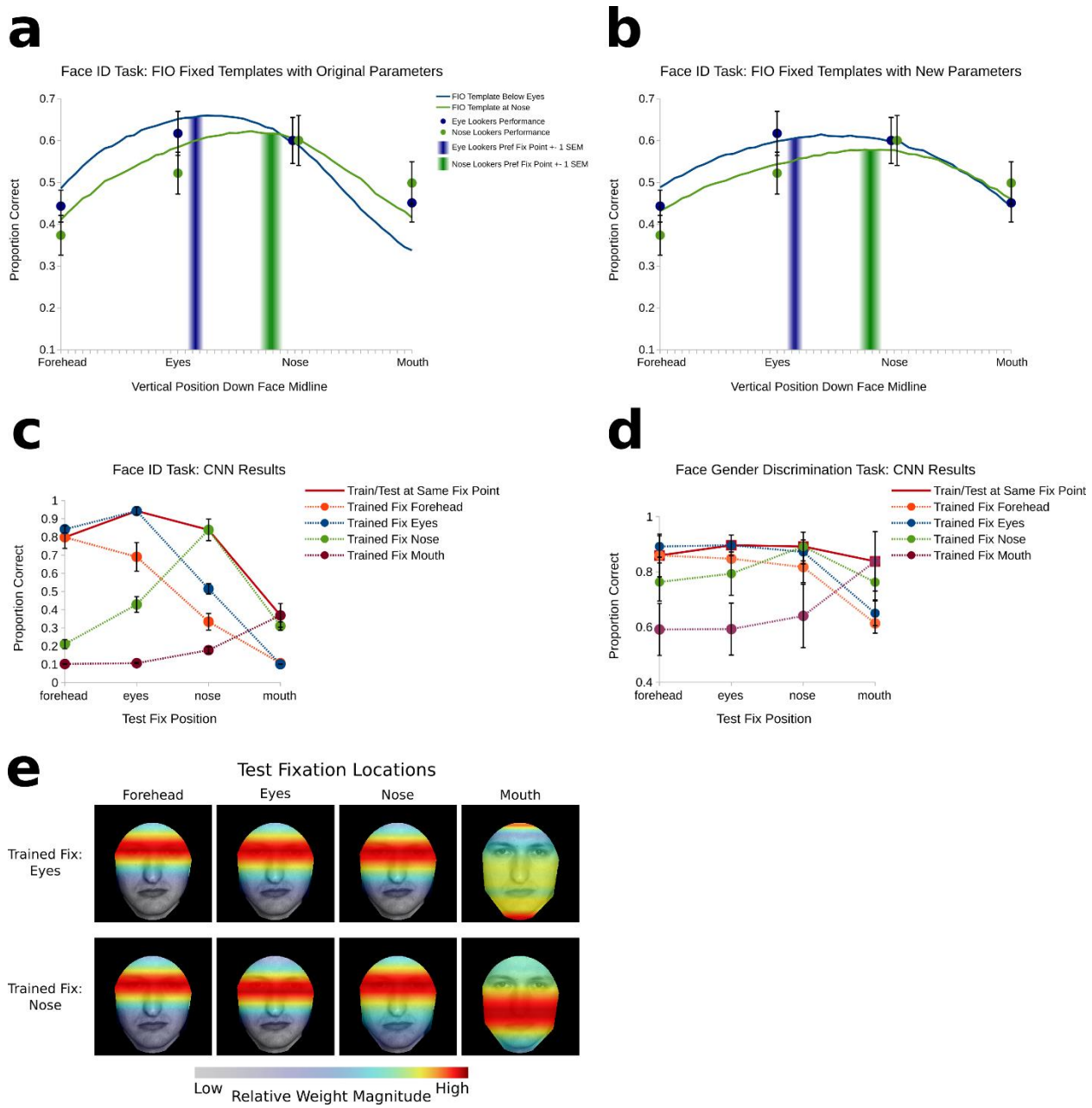

Figure 8: ***a****) The performance profile down the midline of the face for forced-fixation positions at the forehead, eyes, nose, and mouth is shown for eye-lookers (blue points) and nose-lookers (green points). The blue and green bars show the location of the preferred vertical fixation position to a face averaged across eye-lookers and nose-lookers, respectively, with the width of the bars corresponding to one standard error above and below the mean. The performance profile of a modified FIO model with fixed templates is also shown when the fixed template is at a point below the eyes (blue line) and a separate performance profile for when the fixed template is at the nose (green line). Here, the location of the fixed template represents the internal representation of the foveated face stimuli that the model uses to compare an incoming face stimulus to on each simulated trial. In the*

modified FIO model, although the internal representation of the faces is fixed at a specific simulated fixation position, different fixation positions for incoming face stimuli are still tested, which allows us to find a performance profile at different fixation positions. Here, the same parameters are used for the original FIO model, except the internal noise parameters, which is adjusted to best fit the human data for eye-lookers and nose-lookers individually. The FIO model with fixed templates provides a reasonably strong fit to the human data and shows the expected differences in the theoretically optimal points of fixation, which match the location of human preferred points of fixation and empirically optimal points of fixation. ***b****) The same plot is shown as in (a), except now all of the parameters of the FIO with a fixed template position below the eyes are fit to the eye-looker human data to provide a slightly better fit. The same parameters, except for the internal noise, are then used to run the FIO with a fixed template position at the nose. Similar to (a), the FIO with new parameters shows the expected differences in the theoretically optimal points of fixation, which match the location of human preferred points of fixation and empirically optimal points of fixation.* ***c****) The results of a CNN model are shown for a human face identification task, using the same face stimuli as the ones used to run the FIO model and to run the human psychophysics experiments. The differently colored dashed lines represent performance profiles for CNN models separately trained with different stimulus sets where the center of foveation was at the forehead (orange), eyes(blue), nose (green), and mouth (purple). Performance is shown for each of those models being tested with separate stimuli sets that were foveated at the four different fixation positions shown on the x-axis. For all four models, performance is highest at the test fixation position that matches the training fixation position. When there is a mismatch in training and testing fixation positions, there is a steep drop in performance proportional to how far away the training fixation position is form the testing fixation position.* ***d****) The same type of plot is shown as in (c), except for the results of a CNN model trained on a face gender discrimination task instead of a face identification task. We ran this model in order to show that the matched-template principle of performance being maximized when the training fixation position matches the testing fixation position, generalizes to a different face task with different face stimuli. Since gender discrimination is a two-class problem, the performance profiles for each of the trained models are much less steep than for a face identification task, and chance performance is at a proportion correct value of 0.5 instead of at 0.1.* ***e****) A visualization is shown of the parts of the face stimuli that the CNN model trained on a face identification task uses the most. The top and bottom rows represent a model that was trained at a fixation location at the eyes and nose, respectively. The columns represent four test fixation locations for each of the two models shown. The scale used to show the importance of different face features is relative only within each image because each visualization has been normalized such that the features with the highest weights are mapped to the highest values. The visualizations show that the eyes are an important region that the CNN uses to do the classification task, regardless of which fixation position was used in the training stimuli. However, when the test fixation location is very far from the training fixation location, the model makes use of other features besides the eyes, and this varies depending on the training fixation location.*

4.4 Discussion

Previous research has shown that there are individual differences between observers in the preferred location of their first eye movement to a face during a face identification task (M. F. Peterson & Eckstein, 2013), with a large portion of the population fixating just below the eyes ("eye-lookers"), and a gradual drop-off in fixation frequency toward the nose ("nose-lookers"). Within observers, their preferred first eye movement locations have also been shown to be consistent across time and optimal in maximizing performance in a face identification task relative to non-preferred fixation positions. Here, we reproduced fixation behaviors and forced-fixation performance profile differences between eye-lookers and nose-lookers. Then we explored two different theories to explain why these individual differences may occur along with two computational models to help explain the aspects of visual processing that may differ between these two groups.

The first theory that we explored, which we refer to as the "altered-anisotropy theory," involves possible differences in the anisotropy of the retina between eye-lookers and nose-lookers. It is known that differences in the quality of representation of different parts of the visual environment are known to exist between the lower and upper visual field (vertical anisotropy) and have been shown to result in higher performance when stimuli are presented in the lower visual field relative to when the same stimuli are presented in the upper visual field in both simple low-level visual tasks (Marisa Carrasco et al., 2001; Corbett & Carrasco, 2011) as well as in more complex higher-level (Marisa Carrasco et al., 2004; S. He et al., 1996; Intriligator & Cavanagh, 2001; Kristjánsson & Sigurdardottir, 2008) visual tasks. In relation to faces, we investigated if nose-lookers have a ratio of acuity in their upper visual field relative to their lower visual field that is higher than the same ratio in eye-lookers. After

separating observers into groups of eye-lookers and nose-lookers based on averages of preferred vertical initial fixation positions to faces in several face discrimination tasks, we measured the anisotropy of their visual fields with two different tasks: a gabor detection task and a natural image matching task. Similar to previous results with other tasks, we found higher performance, on average, when task-relevant information was presented in the lower visual field relative to the upper visual field, within individual participants. However, we did not find significant differences in the performance ratio comparing visual fields, between groups of eye-lookers and nose-lookers. This provides evidence against the altered-anisotropy theory.

The second theory that we explored, which we refer to as the "matched-template" theory, involves differences in face-specific mechanisms between eye-lookers and nose-lookers, which do not carry over to tasks that do not involve human faces. First, we had participants do a free-fixation chimp face discrimination task, a free-fixation luggage bag discrimination task, and a sports/regular car discrimination task, in order to establish the specificity of eye movements to human faces relative to other stimuli. Although we found significant correlations in vertical fixation position across participants in the human face identification task vs. the chimp face identification task, and in the human face identification task vs. the luggage bag identification task, the correlation in the latter was much lower. The variability in the vertical fixation position in both the chimp face identification task and luggage bag identification tasks was also significantly higher relative to the variability in the human face identification task. In addition, the forced-fixation gender identification task results showed that the FIO model, which was trained on different human data from a face identification task, can generalize well to other human face tasks in the sense of predicting

the forced-fixation performance profile as well as the location of the empirical optimal initial fixation position. This result as well as a generalization of the FIO model to other face tasks has been shown previously (M. F. Peterson & Eckstein, 2012; Tsank & Eckstein, 2017). However, in contrast to other face tasks, the FIO does not generalize well to the forced-fixation data from the sports/regular car discrimination task. The FIO predicts a performance profile where the theoretically optimal point of fixation is located toward the very bottom of the car images, where the important features of the task are. In contrast, the human data shows that humans do not have an empirical optimal initial fixation position to cars for this task and their preferred initial fixation position is far from the theoretical optimal point of fixation predicted by the FIO.

In order to further explore the matched-template theory and how it may explain differences in fixation behavior between eye-lookers and nose-lookers, we simulated differences in face processing based on different fixation-specific representations in the FIO model and separately in a CNN model. In the original FIO model, the internal representation of all face stimuli are dynamic, in terms of the center of the simulation of foveation, and differ based on the current fixation position that is being processed by the model. In contrast to this, we ran two different versions of the FT-FIO model, where the internal representation of a face was fixed by simulating foveation at either a point below the eyes, corresponding to a possible eye-looker representation, or at a point at the nose, corresponding to a possible nose-looker representation. The results of the FT-FIO model predicted the differences in the performance profiles of eye-lookers and nose-lookers well.

In addition to the FIO model, we also ran a CNN model with different inputs of fixation-specific face representations. The FIO model simulates the foveation of the visual

system, but otherwise makes optimal classification decisions and has a perfect pixel-level representation of the stimuli that are being discriminated. In contrast, a CNN model learns its own internal representations of stimuli based on the stimuli sets that are used to train the CNN, which allows it to simulate differences in human visual development. Although they are only a rudimentary approximation of human cortical processing, CNNs are starting to be used in the study of human vision and face processing (see (O'Toole et al., 2018) for a review) after successful implementations of various face classification tasks in computer vision (Li et al., 2015; Schroff et al., 2015; Taigman et al., 2014), some of which have achieved close to human performance. CNNs are known to have certain useful properties that may be able to represent aspects of the human visual system. One of those aspects is a feedforward multilayer structure that represents progressively more complex features starting from edge detection and ending with complex shapes, textures, colors, and the relationships between them. Another important aspect is the ability to learn feature detectors that are adapted to the complex statistical properties of the features in the images that the model is being trained on. Although CNNs are able to learn complex feature representations, the CNN that we used in our simulations does not have an explicit representation of the variable density of photoreceptors and ganglion cells in the human retina, which would allow for the representation of foveation at different fixation positions. Instead, we used different training sets that were "pre-foveated" at four different locations, corresponding to the forehead, eyes, nose, and mouth on the face stimuli. The spatial filtering operations and parameters of the spatially variant contrast sensitivity function (see Methods second for details) that we used to simulate the foveations were the same as those that were used in the FIO. After training the CNN with four different training sets, corresponding to four different fixation-specific face

representations, we then tested the models with all combinations (4x4) of fixation-specific representations. We showed that the CNN was able to qualitatively reproduce the forced-fixation performance profile difference between eye-lookers and nose-lookers. More specifically, the results of the CNN simulations showed that the highest performance for a CNN model trained with a particular fixation-specific training set were achieved with a testing set that was processed in the same fixation-specific manner.

5 Conclusion

5.1 Overview of Eye Movements to Faces

Face perception is an important ability that most humans use multiple times a day. A large amount of information can be extracted from a face in a very short period of time. This crucial information contains both social signals as well as signals about our surroundings. Social information that can be extracted from a face, such as someone's identity, gender, and emotional state, allows us to function efficiently with others. Signals about our environment that attract attention with eye gaze or elicit an emotional response, such as an expression of disgust from noxious stimuli, can also be extracted from a face. These environmental signals from others' faces may act as a warning and allow us to quickly direct our own attention to something important in our surroundings. However, the way in which we extract this information depends on making efficient eye movements to specific task-relevant locations on a face. We make eye movements in order to point the foveal region of our retinas toward objects or features that require the highest resolution of visual processing for particular tasks.

It is known that large amount of the information described above can be extracted with just a single eye movement made to a location on a face in between several important features. The three projects presented in this dissertation dealt with the interaction of the initial eye movement to a face with internal face representations. The first project examined the role of the limitation of having a foveated visual system on configural processing of faces, using a more ecologically valid stimulus set with dynamic facial expressions. The second project examined the effects of natural statistics of facial expressions on the initial

eye movement to a face. The third project examined how individual differences in the initial eye movement to a face may shape the development of internal fixation-specific face representations.

5.2 Contributions to the Face Perception Field

The role of eye movements and configural representations in face processing.

Previous research has shown that performance in human face discrimination tasks can be degraded by manipulating the position of features (i.e. eyes, nose, mouth) within a face stimulus (Civile et al., 2018; Collishaw & Hole, 2000; Tanaka & Farah, 1993). This performance difference is typically attributed to a disruption of face mechanisms in higher-order visual areas of the brain involving feature configurations. In this project, we investigated the possibility of the limitation of foveated processing contributing to this performance difference by performance differences resulting from scrambling face features, into different causes: 1) The proximity of informative features to an optimal point of fixation; 2) Suboptimal fixation strategies; 3) Configural representations in the brain. We used a computational Bayesian model that represents the limitation of a foveated visual system to study changes in the optimal point of fixation across different face configurations. We then use a convolutional neural network (CNN) model, which represents additional limitations of the visual system that may be present in higher-order visual areas, to study configural representations. We found that a CNN model was much better able to represent human performance differences across different face configurations as well as across different fixation positions. We concluded that the vast majority of the magnitude of performance differences across different face configurations may be attributed to configural face mechanisms.

Eye movements during gender discrimination of faces are adapted to the naturally occurring statistics of emotional expressions.

The human visual system programs eye movements for specific tasks by taking into account both the varying resolution of the retina and the distribution of visual task-relevant statistical regularities. Face perception tasks are heavily practiced and involve a very consistent location of important face features, which direct the first eye movement to a performance-maximizing optimal point of fixation below the eyes (M. F. Peterson & Eckstein, 2012). However, it is unknown to what extent humans use even more fine-tuned statistical properties, like facial expression frequencies during specific face discrimination tasks to adapt their initial eye movement accordingly. In this project, we ran a face gender discrimination task with an unusually high frequency of happy expressions (50%), which we measured with an eyewear-embedded camera in the real world to have a much lower incidence of 10%. We showed, using ideal observer analysis and convolutional neural network (CNN) analysis, that there is additional information in the mouth region for happy-expression faces, which leads to an increase in performance and a shift downward of the theoretical optimal point of fixation relative to neutral-expression faces. However, we showed that humans are unable to take advantage of this new information, even when forced to fixate at the new theoretical optimal point of fixation, and do not adjust their initial eye movement. We found that a foveated ideal observer model that has a diminished representation of the mouth and region best predicted the human data. Our results suggest that observers learn an optimal point of fixation to faces using the statistics of occurrence of facial expressions for specific tasks and are inflexible to greatly altered facial expression statistics.

The development of internal fixation-specific face representations.

Previous research has shown that humans have a preferred initial fixation position to faces during common face discrimination tasks. This preferred point is consistent within observers such that they fixate the same point even when tested months apart. However, there are individual differences in the location of this preferred point across observers (M. F. Peterson & Eckstein, 2013). The preferred fixation locations belong to a distribution that shows a majority of observers fixating the eye region, with a continuous and decreasing frequency going down toward the mouth region. In addition, there are differences in observers' empirical optimal points of fixation, such that observes maximize their performance in a face identification task when forced to fixate closest to their individual preferred fixation location. In this project, we divided a set of observers into two groups based on their preferred fixation location: "eye-lookers" and "nose-lookers." We then test two hypotheses that attempted to explain what causes the individual differences between these two groups of observers. The first hypothesis involves differences in contrast sensitivity across the upper and lower visual fields, between eye-lookers and nose-lookers. We tested this hypothesis by measuring the performance of both groups of observers in two lower-level visual tasks that do not involve face stimuli. We ran observers in a forced-fixation condition where they use different parts of their visual fields to do the task. The results showed that there are no significant differences between eye-lookers and nose-lookers when using their lower visual field vs. using their upper visual field to process the stimuli in these tasks. The second hypothesis involved modeling possible differences in internal fixation-specific face representations between eye-lookers and nose-lookers. We

implemented these internal representations in both a foveated ideal observer model as well as a convolutional neural network (CNN) model, and showed that both models were able to represent important differences in human forced-fixation data between eye-lookers and nose-lookers. Our results suggest that individual differences in fixation position between observers are face-specific rather than a more general difference in low-level vision. Our modeling efforts provide evidence that these face-specific differences involve fixation-specific representations in the brain.

Overall Contributions.

Taken together, the results we obtained from these projects added to the growing knowledge that eye movement strategies to face stimuli are a highly practiced and consistent behavior. This behavior can be thought of in the context of the specificity of face processing in the brain, relative to the less specialized processing of other complex objects. We have presented evidence that it is a behavior that depends on the statistics of the faces that humans are exposed to during different face discrimination tasks. In turn, this behavior may also shape the internal representations of faces in our brain, such that those representations are fixation-specific. These results suggest that our internal representations of faces may be closely tied to the way in which our visual system interacts with our visual environment. This allows for an efficient use of our neural resources for the evolutionarily important ability of perceiving faces and extracting meaningful information from them.

6 References

Ackermann, J. F., & Landy, M. S. (2010). Suboptimal Choice of Saccade Endpoint in Search with Unequal Payoffs. *Journal of Vision*, *10*(7), 530–530. https://doi.org/10.1167/10.7.530

Aifanti, N., Papachristou, C., & Delopoulos, A. (2010). The MUG facial expression database. *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, 1–4.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(4), 997–1010.

Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Barlow, H. B. (1980). The absolute efficiency of perceptual decisions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *290*(1038), 71–82.

Barrett, H. H., Yao, J., Rolland, J. P., & Myers, K. J. (1993). Model observers for assessment of image quality. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(21), 9758–9765.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion,

and morphing on scanning fixations. *Perception*, *35*(8), 1089 – 1105. https://doi.org/10.1068/p5547

Belle, G. V., Graef, P. D., Verfaillie, K., Rossion, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation. *Journal of Vision*, *10*(5), 10. https://doi.org/10.1167/10.5.10

Bombari, D., Mast, F. W., & Lobmaier, J. S. (2009). Featural, Configural, and Holistic Face-Processing Strategies Evoke Different Scan Patterns. *Perception*, *38*(10), 1508–1521. https://doi.org/10.1068/p6117

Bukach, C. M., Bub, D. N., Gauthier, I., & Tarr, M. J. (2006). Perceptual expertise effects are not all or none: spatially limited perceptual expertise for faces in a case of prosopagnosia. *Journal of Cognitive Neuroscience*, *18*(1), 48–63. https://doi.org/10.1162/089892906775250094

Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision*, *14*(2). https://doi.org/10.1167/14.2.1

Burge, J., & Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications*, *6*, 7900. https://doi.org/10.1038/ncomms8900

Burge, J., & Jaini, P. (2017). Accuracy Maximization Analysis for Sensory-Perceptual Tasks: Computational Improvements, Filter Robustness, and Coding Advantages for Scaled Additive Noise. *PLOS Computational Biology*, *13*(2), e1005281. https://doi.org/10.1371/journal.pcbi.1005281

Burgess, A. E. (1994). Statistically defined backgrounds: performance of a modified nonprewhitening observer model. *Journal of the Optical Society of America. A,*

Optics, Image Science, and Vision, *11*(4), 1237–1242.

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of Human Visual Signal Discrimination. *Science*, *214*(4516), 93–94.

Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, *42*(8), 949–967. https://doi.org/10.1016/S0042-6989(02)00039-1

Carrasco, MARISA, & Frieder, K. S. (1997). Cortical Magnification Neutralizes the Eccentricity Effect in Visual Search. *Vision Research*, *37*(1), 63–82. https://doi.org/10.1016/S0042-6989(96)00102-2

Carrasco, Marisa, Marie Giordano, A., & McElree, B. (2004). Temporal performance fields: visual and attentional factors. *Vision Research*, *44*(12), 1351–1365. https://doi.org/10.1016/j.visres.2003.11.026

Carrasco, Marisa, Talgar, C. P., & Cameron, E. L. (2001). Characterizing visual performance fields: effects of transient covert attention, spatial frequency, eccentricity, task and set size. *Spatial Vision*, *15*(1), 61–75.

Civile, C., Elchlepp, H., McLaren, R., Galang, C. M., Lavric, A., & McLaren, I. (2018). The effect of scrambling upright and inverted faces on the N170. *Quarterly Journal of Experimental Psychology*, 1747021817744455. https://doi.org/10.1177/1747021817744455

Collishaw, S. M., & Hole, G. J. (2000). Featural and Configurational Processes in the Recognition of Faces of Different Familiarity. *Perception*, *29*(8), 893–909. https://doi.org/10.1068/p2949

Corbett, J. E., & Carrasco, M. (2011). Visual Performance Fields: Frames of Reference.

PLOS ONE, *6*(9), e24470. https://doi.org/10.1371/journal.pone.0024470

Dachille, L. R., Gold, J. M., & James, T. W. (2012). The response of face-selective cortex with single face parts and part combinations. *Neuropsychologia*, *50*(10), 2454–2459. https://doi.org/10.1016/j.neuropsychologia.2012.06.016

de Haas, B., Schwarzkopf, D. S., Alvarez, I., Lawson, R. P., Henriksson, L., Kriegeskorte, N., & Rees, G. (2016). Perception and Processing of Faces in the Human Brain Is Tuned to Typical Feature Locations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(36), 9289–9302. https://doi.org/10.1523/JNEUROSCI.4131-14.2016

Droll, J. A., Abbey, C. K., & Eckstein, M. P. (2009). Learning cue validity through performance feedback. *Journal of Vision*, *9*(2), 18.1-23. https://doi.org/10.1167/9.2.18

Eckstein, M. (2017). Probabilistic Computations for Attention, Eye Movements, and Search. *Annu. Rev. Vis. Sci.*, *3:18.1–18.24*.

Eckstein, M. P., Abbey, C. K., Bochud, F. O., & others. (2000). A practical guide to model observers for visual detection in synthetic and natural noisy images. *Handbook of Medical Imaging*, *1*, 593–628.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. *Current Biology*, *27*(18), 2827-2832.e3. https://doi.org/10.1016/j.cub.2017.07.068

Eckstein, M. P., Schoonveld, W., Zhang, S., Mack, S., & Akbas, E. (2015a). Optimal and human eye movements to clustered low value cues to increase decision rewards during search. *Vision Research*. https://doi.org/10.1016/j.visres.2015.05.016

Eckstein, M. P., Schoonveld, W., Zhang, S., Mack, S. C., & Akbas, E. (2015b). Optimal and human eye movements to clustered low value cues to increase decision rewards during search. *Vision Research, 113*(Pt B), 137–154. https://doi.org/10.1016/j.visres.2015.05.016

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology. Human Perception and Performance, 21*(3), 628–634.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482–498.

Geisler, W. S. (2011). Contributions of Ideal Observer Theory to Vision Research. *Vision Research, 51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Geisler, W. S., & Ringach, D. (2009). Natural Systems Analysis. *Visual Neuroscience, 26*(01), 1–3. https://doi.org/10.1017/S0952523808081005

Gold, J. M., Barker, J. D., Barr, S., Bittner, J. L., Bromfield, W. D., Chu, N., … Srinath, A. (2013). The efficiency of dynamic and static facial expression recognition. *Journal of Vision, 13*(5). https://doi.org/10.1167/13.5.23

Gold, J. M., Mundy, P. J., & Tjan, B. S. (2012). The Perception of a Face Is No More Than the Sum of Its Parts. *Psychological Science*, 0956797611427407. https://doi.org/10.1177/0956797611427407

Golla, H., Ignashchenkova, A., Haarmeier, T., & Thier, P. (2004). Improvement of visual acuity by spatial cueing: a comparative study in human and non-human primates. *Vision Research, 44*(13), 1589–1600. https://doi.org/10.1016/j.visres.2004.01.009

Guo, X. M., Oruç, I., & Barton, J. J. S. (2009). Cross-orientation transfer of adaptation for

facial identity is asymmetric: a study using contrast-based recognition thresholds. *Vision Research*, *49*(18), 2254–2260. https://doi.org/10.1016/j.visres.2009.06.012

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, *34*(3), 479–490.

Haxby, null, Hoffman, null, & Gobbini, null. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology: CB*, *24*(13), R622-628. https://doi.org/10.1016/j.cub.2014.05.020

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv:1512.03385 [Cs]*. Retrieved from http://arxiv.org/abs/1512.03385

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334. https://doi.org/10.1038/383334a0

Heering, A. de, Rossion, B., Turati, C., & Simion, F. (2008). Holistic face processing can be independent of gaze behaviour: Evidence from the composite face illusion. *Journal of Neuropsychology*, *2*(1), 183–195. https://doi.org/10.1348/174866407X251694

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, *33*(1), 98–106. https://doi.org/10.3758/BF03195300

Henriksson, L., Mur, M., & Kriegeskorte, N. (2015). Faciotopy—A face-feature map with face-like topology in the human occipital face area. *Cortex*, *72*, 156–167. https://doi.org/10.1016/j.cortex.2015.06.030

Hidalgo-Sotelo, B., Oliva, A., & Torralba, A. (2005). Human Learning of Contextual Priors for Object Search: Where does the time go? *Computer Vision and Pattern Recognition Workshop*, 86. http://doi.ieeecomputersociety.org/10.1109/CVPR.2005.470

Hills, P. J., Cooper, R. E., & Pake, J. M. (2013). First fixations in face processing: the more diagnostic they are the smaller the face-inversion effect. *Acta Psychologica*, *142*(2), 211–219. https://doi.org/10.1016/j.actpsy.2012.11.013

Hsiao, J. H., & Cottrell, G. (2008). Two Fixations Suffice in Face Recognition. *Psychological Science*, *19*(10), 998–1006. https://doi.org/10.1111/j.1467-9280.2008.02191.x

Intriligator, J., & Cavanagh, P. (2001). The Spatial Resolution of Visual Attention. *Cognitive Psychology*, *43*(3), 171–216. https://doi.org/10.1006/cogp.2001.0755

Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the eye region in neural processing of faces. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(47), 16666–16682. https://doi.org/10.1523/JNEUROSCI.2391-12.2012

Jacques, C., d'Arripe, O., & Rossion, B. (2007). The time course of the inversion effect during individual face discrimination. *Journal of Vision*, *7*(8), 3. https://doi.org/10.1167/7.8.3

Kanade, T., & and, J. F. C. (2000). Comprehensive database for facial expression analysis. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 46–53. https://doi.org/10.1109/AFGR.2000.840611

Kanade, T., Cohn, J. F., & Yingli Tian. (2000). Comprehensive database for facial expression

analysis. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 46–53. https://doi.org/10.1109/AFGR.2000.840611

Kanan, C., Bseiso, D. N. F., Ray, N. A., Hsiao, J. H., & Cottrell, G. W. (2015). Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision Research*, *108*, 67–76. https://doi.org/10.1016/j.visres.2015.01.013

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Koehler, K., & Eckstein, M. (2017). Beyond Scene Gist: Objects guide search more than backgrounds. *Journal of Experimental Psychology. Human Perception and Performance*.

Kristjánsson, A., & Sigurdardottir, H. M. (2008). On the benefits of transient attention across the visual field. *Perception*, *37*(5), 747–764. https://doi.org/10.1068/p5922

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an ideal-observer model of reading. *Psychological Review*, *104*(3), 524–553.

Legge, Gordon E., Hooven, T. A., Klitz, T. S., Stephen Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: new insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234.

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170

Liston, D. B., & Stone, L. S. (2008). Effects of prior information and reward on oculomotor

and perceptual choices. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(51), 13866–13875. https://doi.org/10.1523/JNEUROSCI.3120-08.2008

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8–15. https://doi.org/10.1016/j.tics.2006.11.002

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014a). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision*, *14*(7). https://doi.org/10.1167/14.7.6

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014b). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision*, *14*(7). https://doi.org/10.1167/14.7.6

Morvan, C., & Maloney, L. T. (2012). Human Visual Search Does Not Maximize the Post-Saccadic Probability of Identifying Targets. *PLOS Computational Biology*, *8*(2), e1002342. https://doi.org/10.1371/journal.pcbi.1002342

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391. https://doi.org/10.1038/nature03390

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection

in visual search. *Vision Research*, *49*(10), 1286–1294. https://doi.org/10.1016/j.visres.2008.12.005

Navalpakkam, V., Koch, C., Rangel, A., & Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 5232–5237. https://doi.org/10.1073/pnas.0911972107

Or, C. C.-F., Peterson, M. F., & Eckstein, M. P. (2015). Initial eye movements during face identification are optimal and similar across cultures. *Journal of Vision*, *15*(13), 12. https://doi.org/10.1167/15.13.12

Oruc, I., Shafai, F., Murthy, S., Lages, P., & Ton, T. (2018). The adult face-diet: A naturalistic observation study. *Vision Research*. https://doi.org/10.1016/j.visres.2018.01.001

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

Parr, L. A., Heintz, M., Lonsdorf, E., & Wroblewski, E. (2010). Visual kin recognition in nonhuman primates: (Pan troglodytes and Macaca mulatta): inbreeding avoidance or male distinctiveness? *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *124*(4), 343–350. https://doi.org/10.1037/a0020545

Paulun, V. C., Schütz, A. C., Michel, M. M., Geisler, W. S., & Gegenfurtner, K. R. (2015). Visual search under scotopic lighting conditions. *Vision Research*, *113*(Pt B), 155–168. https://doi.org/10.1016/j.visres.2015.05.004

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and

letter identification. *Vision Research*, *46*(28), 4646–4674. https://doi.org/10.1016/j.visres.2006.04.023

Pérez, P., Gangnet, M., & Blake, A. (2003). Poisson Image Editing. *ACM SIGGRAPH 2003 Papers*, 313–318. https://doi.org/10.1145/1201775.882269

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314–E3323. https://doi.org/10.1073/pnas.1214269109

Peterson, M. F., & Eckstein, M. P. (2013). Individual Differences in Eye Movements During Face Identification Reflect Observer-Specific Optimal Points of Fixation. *Psychological Science*. https://doi.org/10.1177/0956797612471684

Peterson, M. F., & Eckstein, M. P. (2014). Learning optimal eye movements to unusual faces. *Vision Research*, *99*, 57–68. https://doi.org/10.1016/j.visres.2013.11.005

Peterson, M. F., Lin, J., Zaun, I., & Kanwisher, N. (2016). Individual differences in face-looking behavior generalize from the lab to the world. *Journal of Vision*, *16*(7), 12. https://doi.org/10.1167/16.7.12

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics*, *63*(7), 1239–1249.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Beliefs alter holistic face processing ... if response bias is not taken into account. *Journal of Vision*, *11*(13), 17. https://doi.org/10.1167/11.13.17

Richler, J. J., Gauthier, I., Wenger, M. J., & Palmeri, T. J. (2008). Holistic processing of faces: perceptual and decisional components. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(2), 328–342. https://doi.org/10.1037/0278-

7393.34.2.328

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, Mechanisms, and Measures of Holistic Processing. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00553

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica*, *128*(2), 274–289. https://doi.org/10.1016/j.actpsy.2008.02.003

Rutherford, M. D., & Towns, A. M. (2008). Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*(7), 1371–1381. https://doi.org/10.1007/s10803-007-0525-7

Saether, L., Van Belle, W., Laeng, B., Brennen, T., & Øvervoll, M. (2009). Anchoring gaze when categorizing faces' sex: evidence from eye-tracking data. *Vision Research*, *49*(23), 2870–2880. https://doi.org/10.1016/j.visres.2009.09.001

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show Me the Features! Understanding Recognition From the Use of Visual Information. *Psychological Science*, *13*(5), 402–409. https://doi.org/10.1111/1467-9280.00472

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion Leads to Quantitative, Not Qualitative, Changes in Face Processing. *Current Biology*, *14*(5), 391–396. https://doi.org/10.1016/j.cub.2004.02.028

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural

representation. *Annual Review of Neuroscience*, *24*, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Soussignan, R., Chadwick, M., Philip, L., Conty, L., Dezecache, G., & Grèzes, J. (2013). Self-relevance appraisal of gaze direction and dynamic facial expressions: effects on facial electromyographic and autonomic reactions. *Emotion (Washington, D.C.)*, *13*(2), 330–337. https://doi.org/10.1037/a0029892

Stritzke, M., Trommershäuser, J., & Gegenfurtner, K. R. (2009). Effects of salience and reward information during saccadic decisions under risk. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *26*(11), B1-13.

Suchow, J. W., Peterson, J. C., & Griffiths, T. L. (2018). Learning a face space for experiments on human identity. *ArXiv:1805.07653 [Cs]*. Retrieved from http://arxiv.org/abs/1805.07653

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *46*(2), 225–245.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, *35*(21), 3053–3069.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-

295X.113.4.766

Tsank, Y., & Eckstein, M. P. (2017). Domain Specificity of Oculomotor Learning after Changes in Sensory Processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(47), 11469–11484. https://doi.org/10.1523/JNEUROSCI.1208-17.2017

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

Van Belle, G., De Graef, P., Verfaillie, K., Rossion, B., & Lefèvre, P. (2010). Face inversion impairs holistic perception: evidence from gaze-contingent stimulation. *Journal of Vision*, *10*(5), 10. https://doi.org/10.1167/10.5.10

Verghese, P. (2012). Active search for multiple targets is inefficient. *Vision Research*, *74*, 61–71. https://doi.org/10.1016/j.visres.2012.08.008

Walker-Smith, G. J., Gale, A. G., & Findlay, J. M. (1977). Eye movement strategies involved in face perception. *Perception*, *6*(3), 313–326.

Walthew, C., & Gilchrist, I. D. (2006). Target location probability effects in visual search: an effect of sequential dependencies. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(5), 1294–1301. https://doi.org/10.1037/0096-1523.32.5.1294

Xu, B., & Tanaka, J. W. (2013). Does face inversion qualitatively change face processing: an eye movement study using a face change detection task. *Journal of Vision*, *13*(2). https://doi.org/10.1167/13.2.22

Yang, N., Shafai, F., & Oruc, I. (2014). Size determines whether specialized expert processes are engaged for recognition of faces. *Journal of Vision*, *14*(8), 17.

https://doi.org/10.1167/14.8.17

Yeshurun, Y., & Carrasco, M. (1999). Spatial attention improves performance in spatial resolution tasks1Parts of this study were presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology (May 1997) and at the Annual Meeting of the Psychonomics Society (November 1997) and published in Abstract format (Yeshurun and Carrasco, 1997and Carrasco and Yeshurun, 1997, respectively).1. *Vision Research*, *39*(2), 293–306. https://doi.org/10.1016/S0042-6989(98)00114-X

Yin, L., Chen, X., 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*, 1–6. https://doi.org/10.1109/AFGR.2008.4813324

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–759. https://doi.org/10.1068/p160747

Zhang, S., Abbey, C. K., & Eckstein, M. P. (2009). Virtual evolution for visual search in natural images results in behavioral receptive fields with inhibitory surrounds. *Visual Neuroscience*, *26*(1), 93–108. https://doi.org/10.1017/S0952523809090014

Zhang, Y., Pham, B. T., & Eckstein, M. P. (2004). Automated optimization of JPEG 2000 encoder options based on model observer performance for detecting variable signals in X-ray coronary angiograms. *IEEE Transactions on Medical Imaging*, *23*(4), 459–474.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). *Learning Deep*

Features for Discriminative Localization. Retrieved from

https://arxiv.org/abs/1512.04150

7 Appendix

7.1 Appendix: Chapter 2

Bayesian Ideal Observer.

Here, we run several different variants of an ideal observer model, starting with a standard ideal observer, which utilizes image information to achieve the highest possible performance and does not simulate the foveation of the visual system like the FIO described below. We run a face emotion identification task with a set of 60 (20 of the same identities for each of 3 emotions) front-view facial expression movies that are normalized for the position of the eyes and chin as well as for contrast (see the Stimuli subsection of Human Psychophysics Studies above for details). Each face movie for the ideal observer simulations consists of 5 frames, which matches a 200ms presentation time that was used in the forced-fixation condition of Experiment 2 (see Trial Timing section of Experimental Conditions in of Chapter 2 for details). An ideal observer optimally integrates information over time, so for each movie, the 5 frames are concatenated into a single large frame, which effectively treats the time dimension as a spatial dimension. The frames at corresponding times for each face movie now spatially align with frames from the same time period in other movies. On each trial of the simulation, the face movies $\{\mathbf{f}_1,\ldots,\mathbf{f}_{60}\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen. The same contrast and additive white noise that was used for humans is then added to a chosen template, *i*. The input data, $\mathbf{g}$, to the ideal observer on each simulated trial is then the sum of a random (1 of 60) face template, $\mathbf{s}_i$, and external noise,

$\mathbf{n}_{ex}$.

$$\mathbf{g} = \mathbf{s}_i + \mathbf{n}_{ex} \tag{2.1.1}$$

The ideal observer does not have any sources of suboptimality such as internal noise or filtering operations on the face template, $\mathbf{s}_i$, that models foveation. Using Bayes rule, the ideal observer finds a set of posterior probabilities, one for each hypothesis that face f from emotion *e* (happy, said, or afraid) was shown, $H_{e,f}$, given the image data, $\mathbf{g}$. Here we use the index, *f*, to represent a calculated posterior probability for a particular face being shown, in contrast to the index, *i*, which represents the actual ground truth signal that was shown on a particular trial.

The posterior probability, $P(H_{e,f} \mid \mathbf{g})$, is calculated using the prior probabilities, $P(H_{e,f})$, and the likelihood, $P(\mathbf{g} \mid H_{e,f})$, of the image data, $\mathbf{g}$, given the presence of each face, f from emotion *e*:

$$P(H_{e,f} \mid \mathbf{g}) = \frac{P(\mathbf{g} \mid H_{e,f})P(H_{e,f})}{P(\mathbf{g})} \propto P(H_{e,f})P(\mathbf{g} \mid H_{e,f}) = l_f \tag{2.1.2}$$

Then to find the posterior probability, $P(H_e \mid \mathbf{g})$, of the presence of a specific emotion, the sum is found across the posterior probabilities of individual faces belonging to that emotion:

$$P(H_e \mid \mathbf{g}) = \sum_f P(H_{e,f} \mid \mathbf{g}) \tag{2.1.3}$$

The normalizing factor, $P(\mathbf{g})$, in equation (2.1.2) is the same for all posterior probabilities, so it can be ignored without changing the result.

The maximum posterior probability is then chosen as the answer at the end of a simulated trial:

$$decision = \underset{e}{\operatorname{argmax}}(P(H_e \mid \mathbf{g})) \tag{2.1.4}$$

The prior probabilities are uniform constants that are set by the experimenter and do not vary from trial to trial. However, calculating the likelihood requires knowing the statistical distribution of noise that is added to each pixel. On each trial, independent Gaussian noise σ_{ex} is added to each pixel $s_{i,p}$, of a random face template, $\mathbf{s}_i$, where *p* indexes 1 to n (500^2) pixels in the 500x500 image, resulting in a noisy image $\mathbf{g}$. At the pixel level, the likelihood, $l_{e,f,p}$, of an individual pixel, g_p of the data coming from pixel, $s_{e,f,p}$, in face template, *f,* from emotion, *e*, is:

$$l_{e,f,p} = \frac{1}{\sqrt{2\pi\sigma_{ex}^{\ 2}}} \exp\left(-\frac{(g_p - s_{e,f,p})^2}{2\sigma_{ex}^{\ 2}}\right) \tag{2.1.5}$$

As a result of the statistical independence of the image noise, the likelihood, $l_{e,f}$, of the data, $\mathbf{g}$, given the presence of the f^{th} face, from emotion, *e*, can be written as a product of the likelihoods of individual pixels, which reduces to a simpler expression involving the original signal template, $\mathbf{s}_{e,f}$, the image data , $\mathbf{g}$, and the external noise standard deviation σ_{ex}:

$$P(\mathbf{g} \mid H_{e,f}) = P(g_1, \ldots, g_n \mid H_{e,f}) = \prod_p P(g_p \mid H_{e,f})$$

$$= \prod_p \frac{1}{\sqrt{2\pi\sigma_{ex}^{\ 2}}} \exp(-\frac{(g_p - s_{e,f,p})^2}{2\sigma_{ex}^{\ 2}}) \propto \prod_p \exp(\frac{-g_p^{\ 2} + 2g_p s_{e,f,p} - s_{e,f,p}^{\ 2}}{2\sigma_{ex}^{\ 2}}) \qquad (2.1.6)$$

$$l_{e,f} \propto \exp(\frac{-\mathbf{g}^T\mathbf{g} + 2\mathbf{g}^T\mathbf{s}_{e,f} - \mathbf{s}_{e,f}^{\ T}\mathbf{s}_{e,f}}{2\sigma_{ex}^{\ 2}}) \propto \exp(\frac{2\mathbf{g}^T\mathbf{s}_{e,f} - \mathbf{s}_{e,f}^{\ T}\mathbf{s}_{e,f}}{2\sigma_{ex}^{\ 2}})$$

Region of Interest Bayesian Ideal Observer.

In order to understand which regions of a face are important for this particular task we also run a Region of Interest Ideal Observer (ROI), which is a Bayesian Ideal Observer that is separately run using small sections of the face stimuli image at a time. We run the ROI for each frame of all the movies separately (i.e. we simulate the emotion discrimination task using only first frame of each movie, then separately the second frame, and so on) in order to see how discriminative information may change as a facial expression develops The calculations are the same as for the ideal observer, except that in contrast to equation (2.1.1), the data, $\mathbf{g}_s$, is now the sum of a random (1 of 60) face template, $\mathbf{s}_{i,s}$, and external noise, $\mathbf{n}_{ex}$, where *s* indexes the section of the face for which performance is separately calculated:

$$\mathbf{g}_s = \mathbf{s}_{i,s} + \mathbf{n}_{ex} \qquad (2.2.1)$$

The signal $\mathbf{s}_{i,s}$ on each simulated trial is now taken from a specific 30x30 pixel section from a randomly chosen face template, *i*. Figure 2a and 2b show how small sections of a face are processed at a time and likelihoods are found for each section. Figure 2c shows a performance map that is created by sampling different sections across the face stimulus. Here, we run a simulation with 30,000 trials. Due to computational constraints, we only sample the sections every 10th pixel rather than every adjacent pixel, which results in a 47x47 performance map (it is not 50x50 because of the 30px section size). This map is then resized using bilinear interpolation to a 500x500 pixel performance map to match the size of the face images.

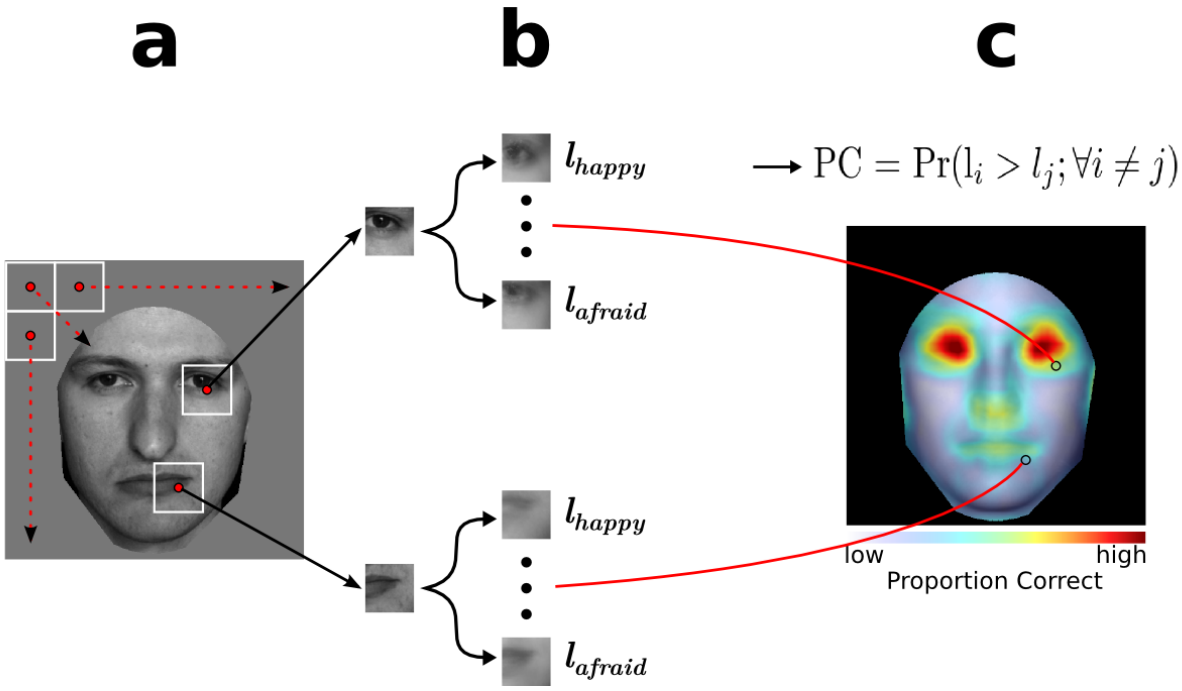

Figure 2: A flow chart for a Region of Interest Ideal Observer. (a) An Ideal Observer is separately run for each small 30x30px section of a face image corresponding to a center point that is sampled every 10px. The ROI is run on a face emotion discrimination task separately on individual frames of each movie (i.e. we simulate the task using only first frame of each movie, then separately the second frame, and so on). (b) On each simulated trial, likelihoods are found for a chosen face to be a particular emotion, which are found from sums of likelihoods of individual identities representing that emotion. (c) The maximum likelihood principle is used to find performance in the task for each separate face section and output a performance map that shows which parts of a face are the most informative for this task.

Foveated Ideal Observer (FIO) Model.

A spatially variant contrast sensitivity function (SVCSF) was used to model the degradation of the quality of information obtained in the periphery of a foveated visual system (M. F. Peterson & Eckstein, 2012):

$$SVCSF(f,r,\theta) = c_0 f^{a_0} \exp(-b_0 f - d_0(\theta) r^{n_0} f) \quad (2.3.1)$$

where *f* is spatial frequency in cycles per degree of visual angle. The terms a_0, b_0, and c_0, were chosen constants set to 1.2, 0.3, and 0.625 respectively, to set the maximum contrast at 1 and the peak at 4 cycles per degree of visual angle at fixation. The polar coordinates *r* and θ specify the distance in visual angle and direction from fixation. d_0 specifies the eccentricity factor as a function of direction, which represents how quickly information is degraded in the periphery. n_0 specifies the steep eccentricity roll off factor. In the model simulations, different parameters are used for d_0 for the vertical up, *du*, vertical down, *dd*, and horizontal, *dh*, directions. The parameters *du*, *dd* , *dh*, and n_0 are fit to the forced-fixation condition in Experiment 2 in order to match human performance (proportion correct) as a function of fixation position (4 different fixations down the vertical midline of the face) of an emotion discrimination task using upright faces. The values used for parameters *du*, *dd*, *dh*, and n_0 respectively, are 2E-6, 9E-6, 1E-6, and 5. The Akaike Information Criterion (Akaike, 1974), which takes into account the variance for each data point, is used as a distance measure. The same parameters are used for the emotion discrimination with all other face configurations (see Stimuli section of Human Psychophysics Studies above). The circular plots between Figure 3b.1 and 3b.2 show examples of 2d contrast

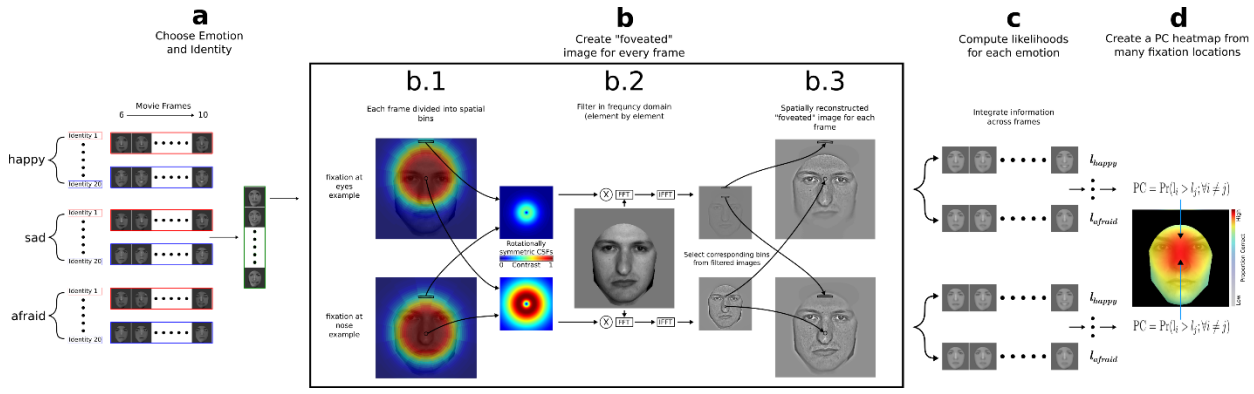

Figure 3: A summary of the process of the computations in the FIO for two fixation positions. The top panels show a fixation point that is below the eyes, which is suboptimal in an emotion discrimination task with upright faces. The bottom panels show a fixation that is above the tip of the nose, which is optimal for this task. ***(a)*** *Many trials are simulated where on each trial, a face template is chosen as a signal. Here, the signal selection is shown for an emotion discrimination task for one of 60 face templates, each of which contains 5 frames (because the model was fit to the short-presentation forced-fixation task of Experiment 2).* ***(b.1-b.3)****, The filtering operation for a noiseless template. The filtering operation is done for each frame separately, after which the frames are concatenated together.* ***(b.1)****, A face image is conceptually divided into bins that correspond to specific CSFs as a function of retinal eccentricity. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the CSF plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial frequencies (seen as a low-contrast blue in the CSF plots).* ***(b.2)****, The image is transformed into the frequency domain, filtered separately by each possible CSF (here only two are shown), and then transformed back into the spatial domain, resulting in a set of differently filtered images corresponding to each bin.* ***(b.3)****, Corresponding bins are then extracted from the filtered images and input into a composite image that simulates foveation. The procedures in b.1–b.3 are then repeated for each of the frames.* ***(c)*** *A set of response variables are then calculated, from which a set of likelihoods is found of each face movie given the noisy image input.* ***(d)****, A decision of which face was shown is made by taking the maximum likelihood. Across many trials, a set of proportion correct (PC) values is found, one for each fixation point, and then combined into a heatmap. iFFT, Inverse FFT.*

sensitivity functions at 2 different locations with respect to the fixation position. Contrast sensitivity functions that correspond to the center of fixation preserve the higher spatial frequencies (seen as a higher contrast in red in the plots), while contrast sensitivity functions that are far from the fixation position act as low-pass filters and mostly leave the low spatial

frequencies (seen as a low contrast in blue in the plots).

Here, we run a face emotion discrimination task using movies of faces (5 frames long) that start with a neutral expression and develop into one of 3 possible expressions that correspond to happiness, sadness, or fear. We separately run several different conditions where the features of the face stimuli are moved or rotated. We simulate many trials of each condition of each task. On each trial of the simulation, the face templates $\{\mathbf{f}_1,\ldots,\mathbf{f}_n\}$ are sampled uniformly at random and a template, $\mathbf{s}_i$, is chosen, where *n* is 60 for the face emotion discrimination task (20 identities, with 3 emotions for each identity). Each face template, $\mathbf{f}_i$, consists of 5 changing frames, which is the length of time that was used for stimulus presentation in the forced-fixation condition of Experiment 2. The 5 frames were taken from frames 6-10 out of 35 frames, in order to account for the average amount of time it took human participants to make the first saccade from the periphery of the screen into the face stimulus. The same contrast and additive white noise that was used for psychophysics experiments in humans is then added to a chosen template, *i*, before being linearly filtered with the SVCSF and corrupted with additional internal white noise to become the input data, $\mathbf{g}_k$, to the ideal observer:

$$\mathbf{g}_k = \mathbf{E}_k(\mathbf{s}_i + \mathbf{n}_{ex}) + \mathbf{n}_{in} \qquad (2.3.2)$$

where $\mathbf{n}_{ex}$ is the external Gaussian white noise, $\mathbf{n}_{in}$ is the internal Gaussian white noise, and $\mathbf{E}_k$ is the linear operator that simulates the fixation dependent foveation of the input. $\mathbf{E}_k$ describes a set of filtering operations, followed by extraction and recombining parts of the filtered noisy templates in the following way: Here, for ease of notation we will describe $\mathbf{E}_k$

as it acts on a random noise free template, $\mathbf{s}_i$, rather than a signal with noise present. However, the computations are the same for a chosen template with added noise. Each combination of eccentricity (r) and direction (θ) from fixation defines its own CSF. The complete set of CSFs can be described in one equation, which we refer to as the SVCSF, where (r) and (θ) remain variables. Due to computational constraints, each image is divided into small bins with a single CSF assigned to each bin. For the emotion discrimination task, each frame $\mathbf{s}_{i,j}$ of each template, $\mathbf{s}_i$, is separately filtered 480 times (30 eccentricities and 16 directions) corresponding to the different CSF functions (and bins) to produce a set of noisy filtered signals. Figure 3b.1 shows an example of two fixations (one at the eyes, and another at the nose) where each face image is conceptually divided into the 480 bins that correspond to different CSF functions relative to the fixation position. Each signal is filtered by taking its fast Fourier transform (FFT), multiplying it on an element by element basis with the corresponding contrast sensitivity function, $\mathbf{CSF}_b$, and then transforming it back into the spatial domain using the inverse FFT (IFFT), resulting in a noisy filtered image $\mathbf{s}_{i,j,b}$, where j represents the frame being processed and b represents the spatial parameters that correspond to bin b (examples of filtered images corresponding to 2 different CSFs are shown between Figure 3b.2 and 3b.3 :

$$\mathbf{s}_{i,b} = IFFT(FFT(\mathbf{s}_i) \circ \mathbf{CSF}_b) \tag{2.3.3}$$

A composite foveated image $\tilde{\mathbf{s}}_{i,k}$, is then formed by extracting the regions of each $\mathbf{s}_{i,b}$ image for the corresponding angle and eccentricity and placing them into $\tilde{\mathbf{s}}_{i,k}$ (Figure 3b.3).

Due to the foveation procedure using the $\mathbf{E}_k$ operator, spatial correlations are formed on the

additive white noise field. In general, it is optimal (maximizes decision accuracy) to utilize templates that undo the spatial correlation through a process known as pre-whitening (Barrett, Yao, Rolland, & Myers, 1993; Burgess, 1994; M. P. Eckstein, Abbey, Bochud, & others, 2000). When using the prewhitening process, correlations are usually corrected by applying various transformations or incorporated into the templates. However, when modeling humans, a common model used is one in which the observer uses templates that match the filtering operations of the visual system (Burgess, 1994; M. F. Peterson & Eckstein, 2012; Y. Zhang, Pham, & Eckstein, 2004). This modeling approach is known as the non-prewhitening with an eye filter (NPWE) and uses templates that match each possible signal with the filtering by the human visual system. This results in a calculation of a set of template responses, $\{r_{1,f,k},\ldots,r_{60,f,k}\}$, where *f* indexes the internal responses to a particular face that was possibly shown during a trial, and *k* indexes different fixation positions to a face image. For a specific fixation position, *k*, and possible input signal, $\mathbf{s}_f$, the template responses are a vector that represents the internal responses of all 60 "face detectors" to that particular input. Here we use the index, *f*, to represent an internal response used to calculate a posterior probability for a particular face being shown, in contrast to the index, *i*, which represents the actual ground truth signal that was shown on a particular trial. Below, we show how to calculate the internal template responses and find the covariance matrix relating them to each other. This is done by first filtering both the original face template, $\mathbf{s}_i$, and the input signal, $\mathbf{s}_f$, plus noise, $\mathbf{s}_f+\mathbf{n}_{ex}$, and then taking the dot product between them as follows:

$$\begin{aligned}
&\mathbf{r}_{f,k}=\{r_{1,f,k},\ldots,r_{60,f,k}\}\\
&=\{(\mathbf{E}_k\mathbf{s}_1)^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in}),\ldots,(\mathbf{E}_k\mathbf{s}_{60})^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})\}\\
&=\{\mathbf{E}_k\mathbf{s}_1,\ldots,\mathbf{E}_k\mathbf{s}_{60}\}^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})
\end{aligned} \tag{2.3.4}$$

Using Bayes rule, the FIO finds a set of posterior probabilities, one for each hypothesis that face *f* from emotion *e* was shown, $H_{e,f}$, given a set of responses $\mathbf{r}_{f,k}$. The posterior probability, $P(H_{e,f} \mid \mathbf{r}_{f,k})$, is calculated using the prior probabilities, $P(H_{e,f})$, and the likelihood, $P(\mathbf{r}_{f,k} \mid H_{e,f})$, of the set of responses given the presence of each face, *f*, and the observer's fixation at spatial location, *k*:

$$P(H_{e,f} \mid \mathbf{r}_{f,k}) = \frac{P(\mathbf{r}_{f,k} \mid H_{e,f})P(H_{e,f})}{P(\mathbf{r}_{f,k})} \propto P(H_{e,f})P(\mathbf{r}_{f,k} \mid H_{e,f}) \qquad (2.3.5)$$

Then to find the posterior probability, $P(H_e \mid \mathbf{r}_{f,k})$, of the presence of a specific emotion, the sum is found across the posterior probabilities of individual faces belonging to that emotion:

$$P(H_e \mid \mathbf{r}_{f,k}) = \sum_f P(H_{e,f} \mid \mathbf{r}_{f,k}) \qquad (2.3.6)$$

The maximum posterior probability is then chosen:

$$decision = \underset{e}{\arg\max}(P(H_e \mid \mathbf{r}_{f,k})) \qquad (2.3.7)$$

The normalizing factor, $P(\mathbf{r}_{f,k})$, in equation (2.3.4) is the same for all posterior probabilities, so it can be ignored without changing the result. The prior probabilities are uniform constants that are set by the experimenter and do not change from trial to trial. However, calculating the likelihood on each trial requires knowing the statistical distribution (means, variances, and covariances) of the template response ($\mathbf{r}_{f,k}$). Noting that $\mathbf{E}_k$ is a linear operator, we will write the distribution of the response, $r_{i,f,k}$, of template, *i*, given face, *f*, using simpler notation. We will denote a single filtered template $\mathbf{s}_{i,k}$, as $\dot{\mathbf{s}}_{i,k}$ and a double filtered template as

$\ddot{\mathbf{s}}_{i,k}$:

$$\begin{aligned} r_{i,f,k} &= (\mathbf{E}_k \mathbf{s}_i)^T (\mathbf{E}_k (\mathbf{s}_f + \mathbf{n}_{ex})) \\ &= (\mathbf{E}_k^2 \mathbf{s}_i)^T (\mathbf{s}_f + \mathbf{n}_{ex}) + (\mathbf{E}_k \mathbf{s}_f)^T \mathbf{n}_{in} \\ &= \ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_f + \ddot{\mathbf{s}}_{i,k}^T \mathbf{n}_{ex} + \dot{\mathbf{s}}_{f,k}^T \mathbf{n}_{in} \end{aligned} \tag{2.3.8}$$

By using zero mean white noise, we are able to write a simple one term expression for the mean of the $r_{i,f,k}$ distribution:

$$\begin{aligned} \mu_{i,f,k} &= E[r_{i,f,k}] = E[\ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_{f,k} + \ddot{\mathbf{s}}_{i,k}^T \mathbf{n}_{ex} + \dot{\mathbf{s}}_{f,k}^T \mathbf{n}_{in}] \\ &= E[\ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_{f,k}] + E[\ddot{\mathbf{s}}_{i,k}^T \mathbf{n}_{ex}] + E[\dot{\mathbf{s}}_{f,k}^T \mathbf{n}_{in}] = \ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_{f,k} \end{aligned} \tag{2.3.9}$$

where $E[\bullet]$ is the expectation operator. The mean of $\mathbf{r}_{f,k}$ when face *f* is chosen is then the vector $\boldsymbol{\mu}_{f,k} = \{\mu_{1,f,k}, \ldots, \mu_{60,f,k}\}$. We are now able to find the covariance between each set of *i*th and *j*th responses that is independent of the presented face, *f*:

$$\begin{aligned} \sum\nolimits_{i,j,k} &= cov(r_{i,f,k}, r_{j,f,k}) = E[(r_{i,f,k} - E[r_{i,f,k}])(r_{j,f,k} - E[r_{j,f,k}])] \\ &= E[(\ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_f + \ddot{\mathbf{s}}_{i,k}^T \mathbf{n}_{ex} + \dot{\mathbf{s}}_{f,k}^T \mathbf{n}_{in} - \ddot{\mathbf{s}}_{i,k}^T \mathbf{s}_f)(\ddot{\mathbf{s}}_{j,k}^T \mathbf{s}_f + \ddot{\mathbf{s}}_{j,k}^T \mathbf{n}_{ex} + \dot{\mathbf{s}}_{f,k}^T \mathbf{n}_{in} - \ddot{\mathbf{s}}_{j,k}^T \mathbf{s}_f)] \\ &= \ddot{\mathbf{s}}_{i,k}^T \ddot{\mathbf{s}}_{j,k} E[\mathbf{n}_{ex}^2] + (\ddot{\mathbf{s}}_{i,k}^T \dot{\mathbf{s}}_{j,k} + \dot{\mathbf{s}}_{i,k}^T \ddot{\mathbf{s}}_{j,k}) E[\mathbf{n}_{ex}^T \mathbf{n}_{in}] + \dot{\mathbf{s}}_{i,k}^T \dot{\mathbf{s}}_{j,k} E[\mathbf{n}_{in}^2] \end{aligned} \tag{2.3.10}$$

Using the property of the expectation of independent random variables, $E[XY] = E[X]E[Y]$, and the fact that $E[\mathbf{n}_{ex}]$ and $E[\mathbf{n}_{in}]$ are both equal to zero, we reduce the middle expression in equation (2.3.9):

$$(\ddot{\mathbf{s}}_{i,k}^T \dot{\mathbf{s}}_{j,k} + \dot{\mathbf{s}}_{i,k}^T \ddot{\mathbf{s}}_{j,k}) E[\mathbf{n}_{ex}^T \mathbf{n}_{in}] = 0 \tag{2.3.11}$$

Then using the property, $Var(X) = E[X^2] - (E[X])^2$, we reduce the first term in equation (2.3.9):

$$\begin{aligned}\ddot{\mathbf{s}}_{i,k}^{T}\ddot{\mathbf{s}}_{j,k}E[\mathbf{n}_{ex}^{2}] &= \ddot{\mathbf{s}}_{i,k}^{T}\ddot{\mathbf{s}}_{j,k}(Var(\mathbf{n}_{ex})+(E[\mathbf{n}_{ex}])^{2}) \\ &= \sigma_{ex}^{2}\ddot{\mathbf{s}}_{i,k}^{T}\ddot{\mathbf{s}}_{j,k}\end{aligned} \tag{2.3.12}$$

where σ_{ex}^{2} is the variance chosen for the external noise. Similarly, the third term in equation 1.9 is reduced:

$$\begin{aligned}\dot{\mathbf{s}}_{i,k}^{T}\dot{\mathbf{s}}_{j,k}E[\mathbf{n}_{in}^{2}] &= \dot{\mathbf{s}}_{i,k}^{T}\dot{\mathbf{s}}_{j,k}(Var(\mathbf{n}_{in})+(E[\mathbf{n}_{in}])^{2}) \\ &= \sigma_{in}^{2}\dot{\mathbf{s}}_{i,k}^{T}\dot{\mathbf{s}}_{j,k}\end{aligned} \tag{2.3.13}$$

This results in a simple expression for the covariance that consists of double filtered templates, single filtered templates, and the variances for the external and internal noise:

$$\sum\nolimits_{i,j,k} = \sigma_{ex}^{2}\ddot{\mathbf{s}}_{i,k}^{T}\ddot{\mathbf{s}}_{j,k} + \sigma_{in}^{2}\dot{\mathbf{s}}_{i,k}^{T}\dot{\mathbf{s}}_{j,k} \tag{2.3.14}$$

Due to the use of Gaussian external noise, the response vector $\mathbf{r}_{f,k}$ for a trial given a face *f* comes from a multivariate normal distribution for which we know the mean and covariance matrix:

$$\mathbf{r}_{f,k} \sim MVN(\boldsymbol{\mu}_{f,k}, \sum\nolimits_{k}) \tag{2.3.15}$$

Knowing $\mathbf{r}_{f,k}$, $\boldsymbol{\mu}_{f,k}$, and $\sum_{k}$ allows us to find the likelihood $l_{f,k}$ of the responses using the multivariate Gaussian probability density function:

$$l_{f,k} = \exp(-\frac{1}{2}(\mathbf{r}_{f,k}-\boldsymbol{\mu}_{f,k})^{T}\sum\nolimits_{k}^{-1}(\mathbf{r}_{f,k}-\boldsymbol{\mu}_{f,k})) \tag{2.3.16}$$

Here, we run a simulation with 100,000 trials. Due to computational constraints, we only run the simulation for fixations corresponding to every 10th pixel, which results in a 50x50 performance map. This map is then resized using bilinear interpolation to a 500x500 pixel performance map to match the size of the face images (Figure 3b.4).

7.2 Appendix: Chapter 3

Ideal Observer.

See equations (2.1.5) and (2.1.6) in the Appendix of Chapter 2 above for details on how class likelihoods are found for an Ideal Observer model.

Foveated Ideal Observer (FIO).

See equations (2.3.8) to (2.3.16) in the Appendix of Chapter 2 above for details of how the means and covariances of a multivariate Gaussian distribution relating template responses for an FIO are calculated to find class likelihoods.

FIO Model With Dampened Features in Internal Face Template.

We run an additional version of the FIO model where we alter the internal representation of face templates that the model uses to compare face signals to on each trial. In this implementation of the FIO, in contrast to equation (2.3.4) in the Appendix of Chapter 1, the calculation of each template response for each possible input signal plus noise, $\mathbf{s}_f + \mathbf{n}_{ex}$, is:

$$\begin{aligned}\mathbf{r}_{f,k} &= \{r_{1,f,k},\ldots,r_{80,f,k}\} \\ &= \{(\mathbf{E}_k\mathbf{d}_1)^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in}),\ldots,(\mathbf{E}_k\mathbf{d}_{80})^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})\} \\ &= \{\mathbf{E}_k\mathbf{d}_1,\ldots,\mathbf{E}_k\mathbf{d}_{80}\}^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})\end{aligned} \qquad (3.1.1)$$

where each $\mathbf{d}_i$ is a dampened version of a face template $\mathbf{s}_i$ that has been altered by lowering the contrast of specific features, such as the mouth or the eyes. This is done by doing an element-by-element multiplication of $\mathbf{s}_i$ with a weight matrix, $\mathbf{w}_{feature}$, which contains values ranging between 0 and 1, with values less than one in the location of the feature whose

contrast is being lowered:

$$\mathbf{d}_i = \mathbf{s}_i \circ \mathbf{w}_{feature} \tag{3.1.2}$$

It is important to note that only the internal representation of the faces are changed, but not the ground truth input signals, $\mathbf{s}_i$, that are chosen on each trial, or each possible input signal plus noise, $\mathbf{s}_f + \mathbf{n}_{ex}$, which the model tries in order to calculate posterior probabilities.

Efficiency Calculation.

It is often useful to assess how well a human performs relative to the upper bound of performance by calculating the absolute efficiency. This is calculated by taking the squared contrast thresholds of humans relative to the ideal observer to achieve a given performance (e.g., experimental accuracy achieved by the human observer (Barlow, 1980; Burgess, Wagner, Jennings, & Barlow, 1981; Pelli, Burns, Farell, & Moore-Page, 2006; Tjan, Braje, Legge, & Kersten, 1995):

$$Efficiency = \frac{C_{ideal}{}^2}{C_{human}{}^2} \tag{3.2.1}$$

For example, if comparing the efficiency between a human observer and an ideal observer, an ideal observer requires much lower signal contrast, C_{ideal} than humans, C_{human}, in order to match the experimentally measured human performance.

In addition, the relative efficiency of the human observer relative to the FIO (M. F. Peterson & Eckstein, 2014) is similarly defined as:

$$Efficiency = \frac{C_{FIO}{}^2}{C_{human}{}^2} \tag{3.2.2}$$

7.3 Appendix: Chapter 4

Ideal Observer.

See equations (2.1.5) and (2.1.6) in the Appendix of Chapter 1 above for details on how class likelihoods are found for an Ideal Observer model.

Foveated Ideal Observer (FIO).

See equations (2.3.8) to (2.3.16) in the Appendix of Chapter 1 above for details of how the means and covariances of a multivariate Gaussian distribution relating template responses for an FIO are calculated to find class likelihoods.

Fixed Template Foveated Ideal Observer (FT-FIO) Model.

In addition to running the FIO model for several different classification tasks, we also run a modified version of the model, which has a fixed internal face template representation. We run it for a face identification task in order to represent possible differences in internal face templates between eye-lookers and nose-lookers. The algorithmic differences between the FIO model and FT-FIO model can be seen in the way that the template responses are calculated relative to equation (2.3.4):

$$\begin{aligned}\mathbf{r}_{f,k} &= \{r_{1,f,k},\ldots,r_{10,f,k}\} \\ &= \{(\mathbf{E}_p\mathbf{s}_1)^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in}),\ldots,(\mathbf{E}_p\mathbf{s}_{10})^T(\mathbf{E}_k(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})\} \\ &= \{\mathbf{E}_p\mathbf{s}_1,\ldots,\mathbf{E}_k\mathbf{s}_{10}\}^T(\mathbf{E}_p(\mathbf{s}_f+\mathbf{n}_{ex})+\mathbf{n}_{in})\end{aligned} \quad (4.1.1)$$

The only difference between equation (2.3.4) and equation (4.1.1) (besides this task involving 10 face exemplars, with one for each of 10 classes, rather than 60 examplars spread across 3 classes) is that the $\mathbf{E}_k$ operator described above, which simulates foveation at a specific fixation position, *k*, now becomes $\mathbf{E}_p$, but only when it is applied to the internal representations of each of 10 face templates $\{\mathbf{s}_1,\ldots,\mathbf{s}_{10}\}$. Here, *p*, denotes the position of a center of fixation that is fixed at a particular point and is independent of the fixation position,

k, which separately determines how a face stimulus, $\mathbf{s}_f$, chosen on a particular trial is foveated at that fixation position. We separately run the FT-FIO model for a fixed position, *p*, at a point below the eyes as well as at a point at the tip of the nose.