Are you from North or South India? A hard face-classification task reveals systematic representational differences between humans and machines

Harish Katti
Centre for Neuroscience, Indian Institute of Science, Bangalore, India

S. P. Arun
Centre for Neuroscience, Indian Institute of Science, Bangalore, India

We make a rich variety of judgments on faces, but the underlying features are poorly understood. Here we describe a challenging geographical-origin classification problem that elucidates feature representations in both humans and machine algorithms. In Experiment 1, we collected a diverse set of 1,647 faces from India labeled with their fine-grained geographical origin (North vs. South India), characterized the categorization performance of 129 human subjects on these faces, and compared this with the performance of machine vision algorithms. Our main finding is that while many machine algorithms achieved an overall performance comparable to that of humans (64%), their error patterns across faces were qualitatively different despite training. To elucidate the face parts used by humans for classification, we trained linear classifiers on overcomplete sets of features derived from each face part. This revealed mouth shape to be the most discriminative part compared to eyes, nose, or external contour. In Experiment 2, we confirmed that humans relied the most on mouth shape for classification using an additional experiment in which subjects classified faces with occluded parts. In Experiment 3, we compared human performance for briefly viewed faces and for inverted faces. Interestingly, human performance on inverted faces was predicted better by computational models compared to upright faces, suggesting that humans use relatively more generic features on inverted faces. Taken together, our results show that studying hard classification tasks can lead to useful insights into both machine and human vision.

Introduction

Humans make a rich variety of judgments on faces, including gender, personality, emotional state, and more. Understanding the underlying features can help endow a variety of artificial-intelligence applications with humanlike performance. While face detection itself has been extensively studied in computer vision (Viola & Jones, 2001; Barnouti, Al-Dabbagh, & Matti, 2016; M. Wang & Deng, 2018), face categorization has largely been studied using only coarse distinctions such as ethnicity (Caucasian/Black/Asian; Brooks & Gwinn, 2010; Fu, He, & Hou, 2014) and gender (Tariq, Hu, & Huang, 2009; Fu, He & Hou, 2014; Y. Wang, Liao, Feng, Xu, & Luo, 2016). Even in humans, only coarse distinctions such as Caucasian/Black have been studied (Brooks & Gwinn, 2010; Fu, He & Hou, 2014). Humans can reliably classify coarse-grained geographical origin in the absence of salient (but potentially informative) cues such as skin color, expressions, cosmetics, ornaments, or attributes such as hair style (Brooks & Gwinn, 2010). Experience-driven biases can contribute to asymmetries in coarse-grained geographical-origin classification (Toole & Natu, 2013).

Despite these advances, several questions remain unanswered. First, what are the underlying features used by humans? Because differences in coarse geographical origins are large, they manifest in a number of face features. This makes it difficult to identify the true subset of features used by humans. Computationally, many successful algorithms—ranging from local binary patterns (Ojala, Pietikäinen, & Harwood, 1994) to deep neural networks (Krizhevsky, Sutskever, & Hinton, 2012)—also use representations that are difficult to interpret. Second, do these algorithms behave as humans do across faces? Answering this question will require both humans and machines to exhibit systematic variations in performance across faces, which is only possible with hard classification tasks. We address both lacunae in this study.

There are broadly two approaches to creating hard classification problems suitable for research. The first is to impoverish the stimuli, thereby making them harder

Citation: Katti, H., & Arun, S. P. (2019). Are you from North or South India? A hard face-classification task reveals systematic representational differences between humans and machines. *Journal of Vision*, *19*(7):1, 1–17, https://doi.org/10.1167/19.7.1.

https://doi.org/10.1167/19.7.1 Received August 12, 2018; published July 1, 2019 ISSN 1534-7362

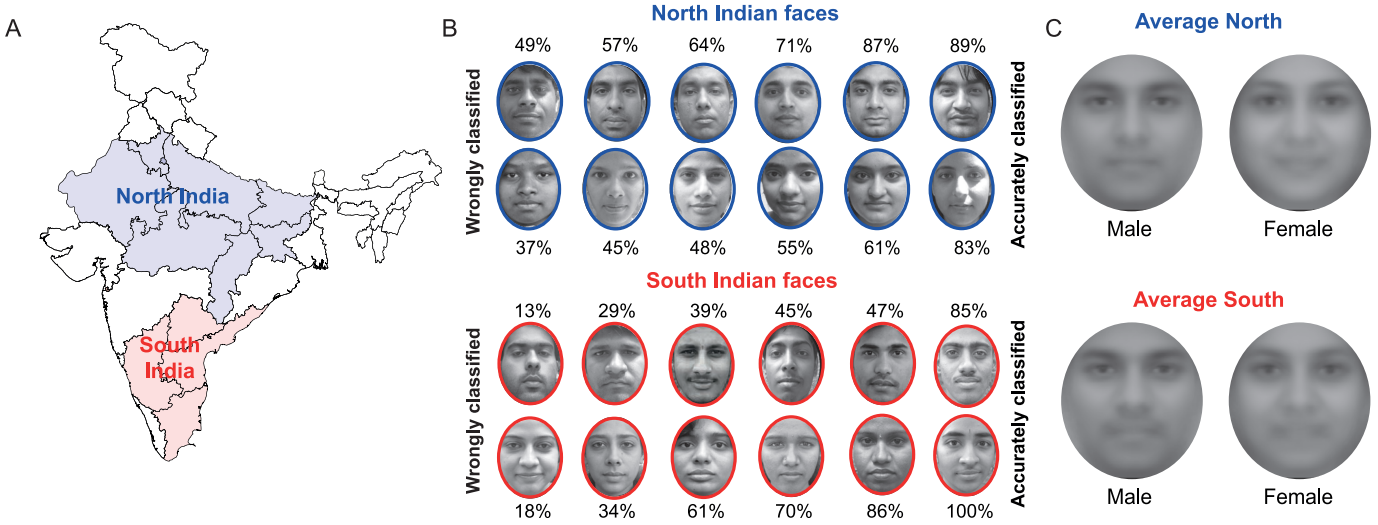

Figure 1. Definitions and examples of North and South Indian faces. (A) Operational definition of northern (blue) and southern (red) regions of India used in this study. We included states that are generally agreed to be part of these regions, and excluded states with unique or distinctive cultural identities (e.g., Kerala, West Bengal, Assam). The fact that independent sets of subjects were able to easily categorize these faces correctly with high accuracy confirms the validity of our definition (see Methods). (B) Example North and South Indian faces. North faces are shown here with a blue border and South faces with a red border, with male faces in the first row and female faces in the second row. Faces are sorted from left to right in ascending order of accuracy with which they were classified correctly across human observers. Each observer saw a given face exactly once. The text above or below each face is the percentage of human observers who correctly classified it into its respective region. In the actual experiment, subjects saw the cropped face against a black background. (C) Average faces. Average male and female North Indian (left) and South Indian (right) faces—obtained by pixel-wise averaging of faces in each group—showing only subtle differences between the faces of the two groups.

to categorize. This can be done by obscuring various portions of a face (Gosselin & Schyns, 2001; Scheirer, Anthony, Nakayama, & Cox, 2014), adding noise (Gold, Bennett, & Sekuler, 1999), binarizing (Kozunov, Nikolaeva, & Stroganova, 2018), or blurring faces (Steinmetz & DaSilva, 2006). Alternatively, the categories themselves can be made more fine-grained, thereby increasing task difficulty. While there has been some work on discriminating between finer grained geographical origin, such as with Chinese/Japanese/Korean (Y. Wang et al., 2016), Chinese subethnicities (Duan et al., 2010), and Myanmar (Tin & Sein, 2011), these studies have not systematically characterized human performance. In fact, it is an open question whether and how well humans can discriminate fine-grained face attributes across various world populations.

Here we present a fine-grained face-classification problem on Indian faces that involves distinguishing between faces originating from northern or southern India. India contains over 12% of the world's population, with large cultural variability. Its geography can be divided roughly into northern and southern regions (Figure 1A) that have stereotyped appearances and social and cultural identities with strong regional mixing. This is accompanied by stereotyped face structure (Figure 1B). Many Indians are able to classify other Indian faces as belonging to specific regions or even states in India but are often unable to describe the face features they are using to do so. Our goal was therefore to characterize human performance on this fine-grained face-classification task and elucidate the underlying features using computational models. Furthermore, our data set, IISCIFD, which we are making publicly available (https://github.com/harish2006/IISCIFD), adds to the relatively few data sets available for Indian faces (Somanath, Rohith, & Kambhamettu, 2011; Setty et al., 2013; Sharma & Patterh, 2015).

Overview

We performed three experiments. In Experiment 1, we created a large data set of North and South Indian faces and characterized the categorization performance of 129 human subjects. We then trained and analyzed computational models to identify face features that predict human performance. We found that computational models trained on face classification showed qualitatively different classification compared to humans. Further, mouth-shape features contributed the most toward predicting human classification. In Experiment 2, we confirmed that humans indeed rely on mouth shape by comparing their classification on faces with occluded parts. In Experiment 3, we investigated whether subjects made qualitatively different responses

when they were shown only briefly flashed or inverted faces.

Experiment 1: Fine-grained face classification by humans

In Experiment 1, we created a large data set of faces and characterized human classification on these faces. We also trained computational models on both geographical-origin labels and human accuracy. To identify the face parts that contribute the most toward classification, we asked how well models trained on overcomplete representations of each face part can predict human accuracy.

Methods

Data set: Our operational definition for North and South Indians is illustrated in Figure 1A. We included states that are representative of North and South India and excluded states with unique or ambiguous identity. The fact that our subjects were easily able to use this classification confirms the validity of our definition. Our face data set has a total of 1,647 Indian faces drawn from two sets of faces, as summarized in Table 1. We refer to this combined dataset as the IISCIFD.

Set 1 consists of 459 face images collected with informed consent from volunteers in accordance with a protocol approved by the Institutional Human Ethics Committee of the Indian Institute of Science. Volunteers were photographed in high resolution (3,648 × 2,736 pixels) against a neutral background. Photographs were collected primarily from volunteers who declared that they as well as both parents belong to a North Indian or South Indian state. For exploratory purposes, we also included the faces of 110 volunteers who declared themselves to be from other regions in India (e.g., Kerala, West Bengal). In addition to their geographical origin, volunteers were requested to report their age, height, and weight.

Set 2 consists of 1,188 faces selected from the Internet after careful validation. Since Indian names are strongly determined by ethnicity, we first identified a total of 128 typical first and 325 last names from each region based on independently confirming these choices with four other Indian colleagues (who were not involved in subsequent experiments). Example first names were Birender and Payal for North India, Jayamma and Thendral for South India. Example last names were Khushwaha & Yadav for North India, Reddy and Iyer for South India. We then used Google Image application programming interfaces to search for face photographs associated with combinations of these typical first and last names. Frontal faces were detected using the CART face detector provided in MATLAB's (MathWorks, Natick, MA) Computer Vision Toolbox, and faces in high resolution (at least 150 × 150 pixels) for which at least three of four colleagues (same as those consulted for names) agreed upon the geographical-origin label were included. These faces were then annotated for gender as well.

Face set	Total	Male	Female	North	South	Other
Set 1	459	260	199	140	209	110
Set 2	1,188	710	478	636	552	0
Total	1,647	970	677	776	761	110

Table 1. Summary of Indian face data set. Set 1 consisted of face photographs taken with consent from volunteers who declared their own geographical origin. Set 2 consisted of face images downloaded from the Internet.

Validation of Set 2: Because Set 2 faces were sourced from the Internet, we were concerned about the validity of the geographical-origin labels. We performed several analyses to investigate this issue. For this and all following statistical comparisons, we first performed the Anderson–Darling test (Anderson & Darling, 1952) to assess normality, and then used either parametric or nonparametric tests as applicable. First, post hoc analysis of classification accuracy revealed that human accuracy on Set 2 (63.6%) was similar to that on Set 1 (62.88%), and this difference was not statistically significant ($p = 0.51$, rank-sum test comparing response-correct labels of faces in the two sets). Second, we asked whether human performance was similarly consistent on the two sets. To this end, we randomly selected responses of 20 subjects from each set and calculated the correlation between the accuracy of two halves of subjects. We obtained similar correlations for the two sets (Set 1: $r = 0.73 \pm 0.05$; Set 2: $r = 0.71 \pm 0.02$; correlation in Set 1 > Set 2 in 585 of 1,000 random subsets). Finally, we asked whether classifiers trained on Set 1 and Set 2 generalized equally well to the other set. For instance, it could be that the labels of Set 2 were noisier and therefore constituted poorer training data. To this end, we selected 400 faces from each set and trained a linear classifier based on spatial and intensity features on geographical-origin classification. The classifier trained on Set 1 achieved an accuracy of 66.4% on Set 1 and generalized to Set 2 faces with an accuracy of 55.2%. Likewise, the classifier trained on Set 2 achieved an accuracy of 61% on Set 2 and generalized to Set 1 with an accuracy of 56.5%. Thus, classifiers trained on either set generalized equally well to the other set. In sum, the overall accuracy and consistency of human subjects, as well as feature-based classification accuracy, were all extremely similar on both sets. Based on these analyses we

combined the geographical-origin labels of both sets for all subsequent analyses.

Image preprocessing: We normalized each face by registering it to 76 facial landmarks (Milborrow & Nicolls, 2014), followed by rotation and scaling such that the midpoint between the eyes coincided across faces and the vertical distance from chin to eyebrow became 250 pixels without altering the aspect ratio. We normalized the low-level intensity information across faces in the data set, since some photographs were taken outdoors, using histogram equalization (function *histeq* in MATLAB) to match the intensity distribution of all faces to a reference face in the data set.

Human behavior

Subjects: A total of 129 subjects (52 women, 77 men; aged 18–55 years) with normal or corrected-to-normal vision performed a binary, face-based geographical-origin classification task. All experimental procedures were in accordance with a protocol approved by the Institutional Human Ethics Committee of the Indian Institute of Science, Bangalore.

Task: Subjects were first introduced to our working definition of North and South Indian regions and were asked to inspect a set of 10 North and 10 South Indian faces (with equal numbers of male and female faces) that were not used for the subsequent classification task. They then performed a classification task consisting of several hundred trials. On each trial, a salt-and-pepper noise mask appeared for 0.5 s, followed by a fixation cross for 0.5 s. This was followed by a face shown for 5 s or until a response was made. Trials were repeated after a random number of other trials if a response was not made within 5 s; we found post hoc that such repeats were indeed very rare (less than 1% of total trials) and did not occur at all for most subjects. Subjects were instructed to indicate using a key press (N for North, S for South) whether the face shown was from North or South India. They were instructed to be fast and accurate, and no feedback was given about their performance. Subjects were allowed to pause and resume the experiment using appropriate key presses to avoid fatigue. Each face was shown only once to a given subject, and a given subject saw on average 259 faces. The average number of subjects per face was 41 for Set 1 and 28 for Set 2.

Computational models

To elucidate the features used by humans for face classification, we compared human performance with that of several computational models. We selected popular models from the computer-vision literature: local binary patterns (LBPs), histograms of oriented gradients (HOGs), and deep convolutional neural networks (CNNs). We also evaluated the performance of simple spatial and intensity features extracted from each face. However, the problem with these models is that their underlying features are difficult to tease apart. Therefore, to elucidate the contribution of individual face parts to human performance, we evaluated the performance of a number of part-based models based on features extracted from specific face parts.

Local Binary Patterns (LBP) and Histogram Oriented Gradients (HOG): We extracted LBP features over tiled rectangular 3×3, 5×5, and 7×7 patches and obtained a 1,328-dimensional feature vector for each face. Our approach is similar to that of Ahonen, Hadid, and Pietikainen (2006). HOG features over eight orientations were extracted over similar patches as LBP, and we obtained a dense 6,723-dimensional HOG feature vector for each face. Our approach is also similar in spirit to that of Déniz Bueno, Salido, & De la Torre (2011).

CNN models (CNN-A, CNN-G, CNN-F): The first CNN, VGG-Face (Parkhi, Vedaldi, & Zisserman, 2015), is a face-recognition CNN which we refer to as CNN-F. The second is a CNN trained for age classification (Levi & Hassner, 2015), which we refer to as CNN-A. The third is a CNN trained for gender classification (Levi & Hassner, 2015), which we refer to as CNN-G. CNN-A and CNN-G consist of three convolutional layers with respective filter sizes of $96 \times 7 \times 7$, $256 \times 5 \times 5$, and $384 \times 3 \times 3$, followed by two 512-node fully connected layers and a single-node decision layer. CNN-F, on the other hand, is a much deeper network and has 11 convolutional layers with filter sizes varying from $64 \times 3 \times 3$ to $512 \times 7 \times 7$, five max pool layers, and three fully connected layers (Parkhi et al., 2015). We used the penultimate 512-dimensional feature vector for each face from CNN-A and CNN-G and a 4,069-dimensional feature vector from CNN-F.

Spatial and intensity features (S, I, SI, IP, SIex, Mom): We also compared computational models based on spatial and intensity features extracted from each face. The spatial features were obtained by measuring a number of 2-D distances between various face parts of interest, and intensity measurements which are based on statistics of intensity in each local region of the face. We tested two approaches to evaluate these features: selective sampling and exhaustive sampling of features.

First, we selectively sampled spatial distances between specific landmarks and sampled intensity statistics within specific regions in the face. We started by registering an active appearance model (Milborrow & Nicolls, 2014) to each face in order to identify 76 facial landmarks, as illustrated in Figure 2A. These landmarks were then used to delineate patches, and mean, minimum, and maximum intensity values were recorded along with landmark-based spatial features, yielding

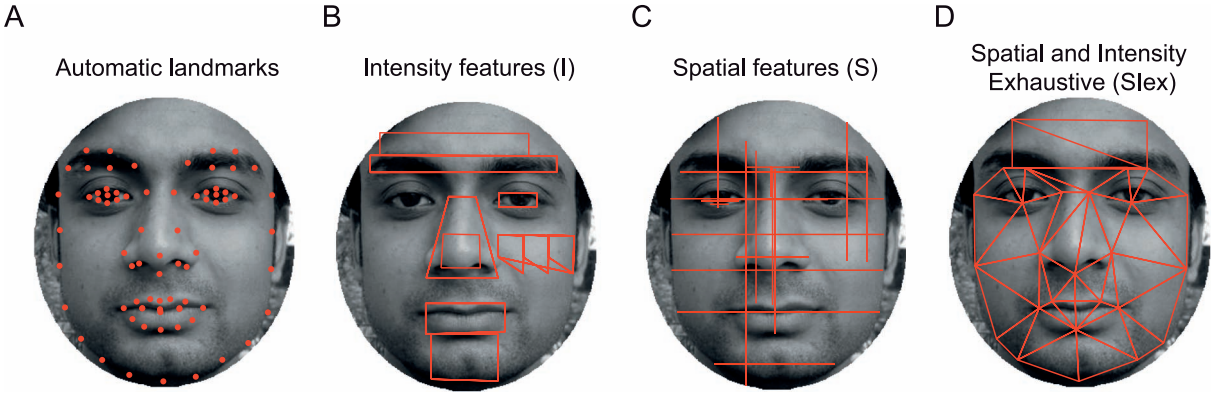

Figure 2. Spatial and intensity feature extraction from faces. (A) Each face was registered to 76 active-appearance-model landmarks using standard methods. (B) We defined a basic set of intensity features (I) from manually defined regions on the face. (C) We defined a basic set of spatial features (S), which were manually defined distances between specific landmarks on the face. (D) We also defined a more exhaustive set of spatial and intensity features by calculating all pair-wise distances between a subset of landmarks (for spatial features) and intensity features from regions demarcated using Delaunay triangulation on these landmarks. This exhaustive set of features is denoted in subsequent analyses as SIex.

a set of 23 spatial (S) and 31 intensity (I) measurements (Figure 2B).

Second, we exhaustively sampled all possible pairs of 2-D distances and intensity measurements. We employed Delaunay triangulation (Delaunay, 1934) over a restricted set of 26 landmarks from which we extracted 43 face patches (Figure 2C), each of which covered the same region across all subjects. We extracted 325 pairwise distances from these 26 landmarks and additionally extracted the mean, minimum, and maximum intensities on all 43 patches, yielding 129 intensity measurements. Together these features are referred to as SIex. To investigate the possibility that global intensity statistics may also contribute to classification, we included the first six moments of the pixel intensity distribution (Mom).

Local face features (E, N, M, C, Eb, IP, ENMC): To model local shape, we selected the following face features: eyes (E), nose (N), mouth (M), contour (C). We also evaluated the performance of all these features concatenated together, which we refer to as ENMC. In addition we included eyebrow (Eb) shape as well.

In each case, we modeled local shape by calculating all pair-wise distances across landmarks related to the two eyes (9C2 for each eye = 72 distances), nose (12C2 = 66), mouth (18C2 = 153), eyebrows (15C2 = 105) and face contour (15C2 = 105). We also calculated 7C2 = 21 configural features by taking the centroid-to-centroid inter-part (IP) distances between all 21 pairs of seven parts which consisted of left eye, right eye, nose, mouth, left-contour, right-contour and chin region.

Model training and cross validation

For each model, we reduced feature dimensionality by projecting feature vectors corresponding to each face along their principal components and retained projections that explain 95% of the variance in the data (Katti, Peelen, & Arun, 2017). Models for binary geographical-origin classification and gender classification were trained using linear discriminant analysis implemented in the MATLAB classify function. Regression models to predict age, height, and weight were trained using regularized linear regression implemented in the MATLAB *lasso* function, which optimizes the squared error subject to a sparsity constraint. We performed 10-fold cross validation throughout to avoid overfitting. In all cases, model performance is reported by concatenating the predictions across the 10 folds and then calculating the correlation with observed data.

Results

We generated a data set with 1,647 Indian faces labeled with fine-grained geographical origin and gender. A fraction of faces also contained self-reported age ($n = 459$), height ($n = 218$), and weight ($n = 253$). We obtained fine-grained geographical-origin classification performance from total of 129 subjects. We tested a number of computational models for their ability to predict all these data given the face image.

Our results are organized as follows. First, we describe human performance on fine-grained geographical-origin categorization. Second, we evaluate the ability of computational models in predicting the same labels. We found that while several models achieved human levels of performance, their error patterns were qualitatively different. Third, we evaluate whether models can predict human accuracy when

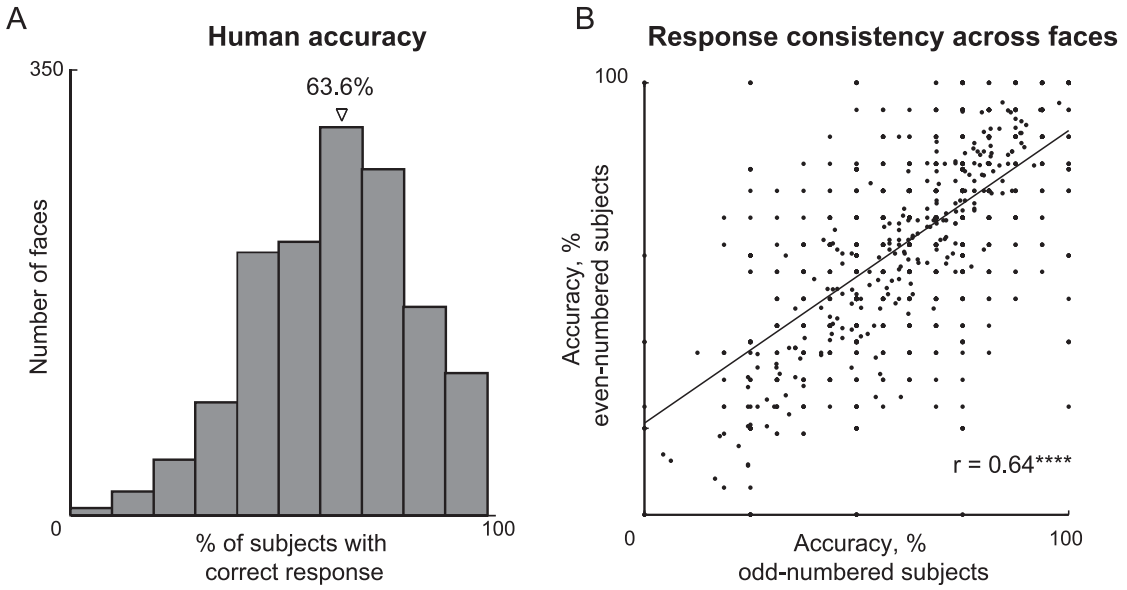

Figure 3. Human performance on fine-grained face classification. (A) Distribution of human accuracy across faces on the North/South Indian face-classification task. Accuracy is calculated as the percentage of participants who correctly guessed the label. (B) Human accuracy for each face calculated from even-numbered subjects plotted against that obtained from odd-numbered subjects. The high correlation indicates that humans were highly consistent: Faces that were accurately judged by one group of subjects were also accurately judged by another independent group.

explicitly trained on this data. This yielded improved predictions but models were still far from human performance. By comparing features derived from individual face parts, we were able to elucidate the face parts that contribute to human accuracy. Fourth, we investigate whether these models can predict other associated labels such as gender, age, height, and weight.

Human performance

In all, we obtained responses from 129 subjects for 1,423 faces across both sets, with over 16 responses for each face. Example faces are shown in Figure 1. Subjects found the task challenging: The average accuracy was 63.6%, but this performance was significantly above chance ($p < 0.0005$; sign-rank test comparing response-correct labels across 1,423 faces against a median of 0.5). Nonetheless, there were variations in accuracy across faces, as shown in Figure 3A. These variations were highly systematic, as evidenced by a high correlation between the accuracy obtained from one half of subjects with that of the other half ($r = 0.64$, $p < 0.0005$; Figure 3B).

In addition to measuring categorization performance, we collected information from each subject about their own geographical origin. A small number of subjects tested were of non-Indian origin ($n = 6$). Their performance was essentially at chance (average accuracy = 51%), and so we removed their data from all further analysis. The remaining 129 subjects were chosen for further analyses.

To investigate whether subjects' own geographical origin affected their performance, we compared the average accuracy of each subject on faces matching their own geographical origin with accuracy on faces from a different geographical origin. This revealed no systematic difference (mean accuracy ± *SD*: 63.8% ± 13.4% for own, 64.5% ± 12.2% for other; $p < 0.55$; rank-sum test). Likewise, subjects showed no preferential bias for their own or other gender in their classification performance (mean accuracy ± *SD*: 65.1% ± 3.2% for own, 63.4% ± 3.1% for other; $p = 0.21$; rank-sum test).

Predicting fine-grained geographical-origin labels

We then evaluated the ability of computational models to predict the ground-truth geographical-origin labels. The cross-validated performance of all the models is summarized in Table 2. Three models yielded equivalent accuracy (i.e., 63% correct): spatial and intensity features (SI), HOG, and CNN-F. To evaluate how local features contribute to geographical origin, we calculated pair-wise spatial distances between facial-landmark points on each specific part of the face (eye, nose, mouth, eyebrows, and contour). This yielded an extensive set of measurements for each part that contained a complete representation of its shape. We then asked which of these feature sets (or a combination thereof) are most informative for classification.

	N/S classification accuracy			Correlation with human accuracy		
Model	*df*	%	Rank	*df*	Correlation	Rank
#Faces	-	1,537	-	-	1,423	-
Human	-	64% ± 6.7%	-	-	0.76	-
S	10	54% ± 0%*	11	11	0.19 ± 0.00*	10
I	**14**	**63% ± 0%**	**2**	15	0.33 ± 0.00*	2
SI	**24**	**63% ± 0%**	**1**	**24**	**0.36 ± 0.00**	**1**
SIex	56	57% ± 1%*	8	57	0.24 ± 0.01*	6
Mom	2	50% ± 0%*	17	2	0.16 ± 0.00*	12
LBP	172	54% ± 0%*	12	157	0.12 ± 0.00*	18
HOG	**487**	**63% ± 1%**	**3**	423	0.29 ± 0.01*	4
CNN-A	124	59% ± 0%*	6	121	0.29 ± 0.01*	3
CNN-G	53	58% ± 1%*	7	50	0.22 ± 0.00*	7
CNN-F	737	61% ± 1%*	4	722	0.20 ± 0.01*	9
E	5	51% ± 1%*	14	5	0.14 ± 0.01*	16
N	7	53% ± 1%*	13	7	0.15 ± 0.00*	13
M	6	56% ± 0%*	9	6	0.20 ± 0.00*	8
C	5	51% ± 1%*	16	5	0.15 ± 0.00*	14
Eb	4	51% ± 1%*	15	5	0.13 ± 0.00*	17
IP	6	49% ± 1%*	18	6	0.14 ± 0.01*	15
ENMC	16	56% ± 1%*	10	16	0.18 ± 0.00*	11

Table 2. Model performance on geographical-origin classification. We trained each model on the ground-truth geographical-origin labels (North vs. South). The numbers in the table under % indicate the mean ± standard deviation of 10-fold cross-validated accuracy of each model across 100 splits. *Asterisks indicate that the model's performance is below the best model in more than 95 of the 100 cross-validated splits, which we deemed statistically significant. Rows in boldface correspond to the best models for either overall accuracy or human accuracy. *Notes*: *df* = degrees of freedom/number of principal components; S = spatial features; I = intensity features; SI = spatial and intensity features; SIex = exhaustive spatial and intensity features; Mom = moments of face-pixel intensity; LBP = local binary patterns; HOG = histogram of oriented gradients; CNN-A, CNN-G, CNN-F = deep networks; E = eye; N = nose; M = mouth; C = contour; Eb = eyebrows; IP = interpart distances; ENMC = eye, nose, mouth, and contour together (see text).

The results are summarized in Table 2. Mouth shape was the most discriminative part for this classification task, and including all other face parts did not improve performance.

Comparing machine predictions with human classification

Next, we wondered whether faces that were easily classified by humans would also be easy for the models trained on ground-truth labels to classify. This would indicate whether humans and computational models use similar feature representations. To this end, we computed the correlation of accuracy/error patterns between every pair of models as well as between human accuracy/errors and all models. To compare these correlations with human performance, we calculated the correlation between the average accuracy of two halves of human subjects. However, this split-half correlation underestimates the true reliability of the data, since it is derived from comparing two halves of the data rather than the full data. We therefore applied a Spearman–Brown correction (Spearman, 1910; Brown, 1910) on this split-half correlation to estimate the true reliability of the data, which is given as $rc = 2r/(r+1)$, where rc is the corrected correlation and r is the split-half correlation.

In the resulting color map, shown in Figure 4, models with similar error patterns across faces show high correlations. Importantly, error patterns of all models were poorly correlated with human performance (Table 2, Figure 4B). This poor correlation between model and human errors could result potentially from models being trained on a mix of weak and strong geographical-origin labels, or from different feature representations. To distinguish between these possibilities, we trained models directly to predict human accuracy using regression methods. Since different features could contribute to an accurately classified North face and a South face, we trained separate models for each class and then concatenated their predictions. The resulting model performance is summarized in Figure 4C. Despite being explicitly trained on human performance, models fared poorly in predicting it.

We conclude that human performance cannot be predicted by most computational models, which is indicative of different underlying feature representations.

Finally, we asked whether the agreement between the responses of two different humans was in general better than the agreement between two models using different feature types. We performed this analysis on faces from Set 1 that had a higher number of human responses than Set 2. The average correlation between correct response patterns for two human subjects ($r = 0.46$) was higher than the average pair-wise correlation between models trained for North/South categorization ($r = 0.08$). The average correlation between two human subjects was also significantly higher ($p < 0.0001$, rank-sum test) than the agreement between models trained to predict human accuracy ($r = 0.35$).

Face-part information in model representations

Since mouth shape was most informative for North vs. South geographical-origin labels (Table 2), we asked whether models trained on whole faces were preferentially encoding mouth information. To this end, we calculated the correlation in the accuracy across faces for each whole-face model with the accuracy predicted by models based on eyes, nose, mouth, or eyebrow

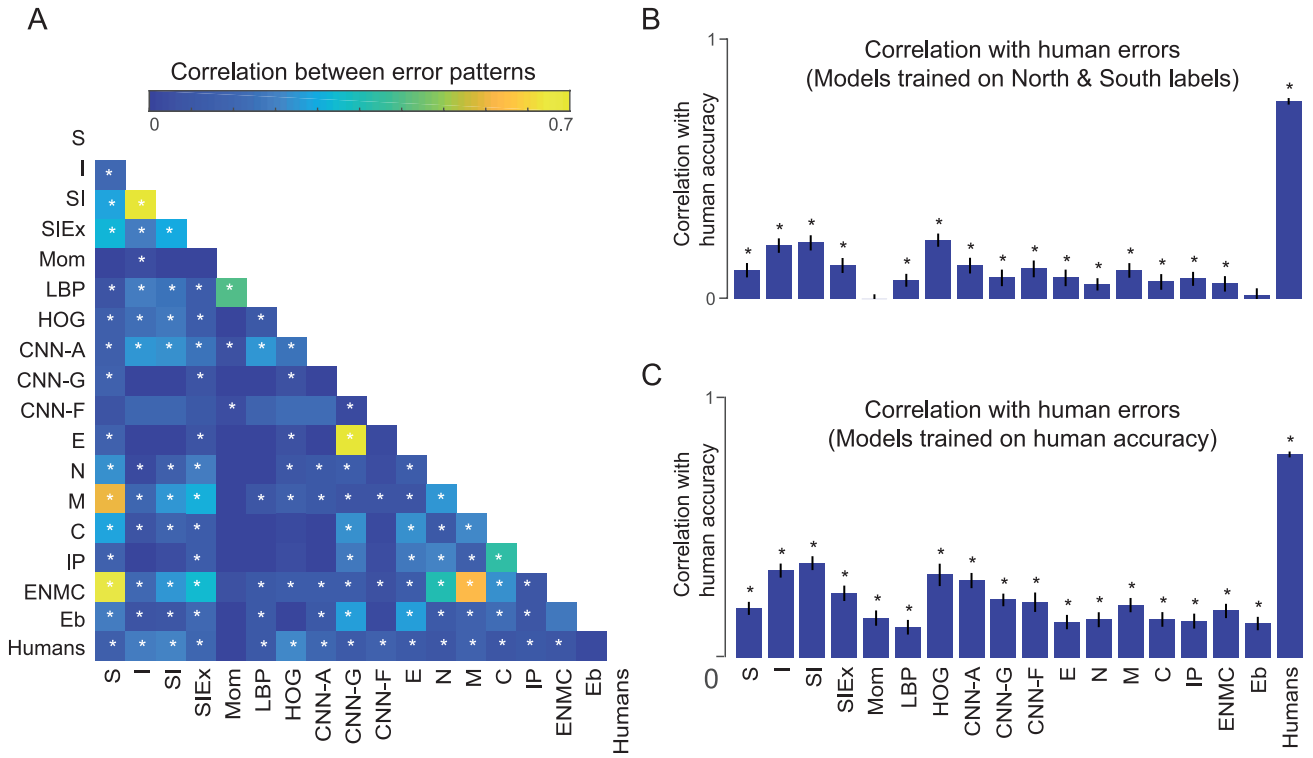

Figure 4. Comparison of error patterns between models and humans. (A) Significant pair-wise correlations between the accuracy and error rates across faces for models and humans ($p < 0.05$). We deemed a correlation significant if less than 5% out of 100 cross-validated splits were below zero. A high artifactual correlation ($r = 1$) between CNN-G and eye shape is marked with #. (B) Correlation between the accuracy of each model (trained on ground-truth labels) and human accuracy across faces. Error bars represent the standard deviation calculated across 1,000 iterations in which faces were sampled with replacement. The rightmost bar depicts human reliability—that is, correlation between average accuracy of one half of subjects with that of the other half of subjects. Significance is calculated as in (A). (C) Correlation between predicted and observed average human accuracy for each model. Here, models were trained to predict human accuracy. Significance is calculated as in (A).

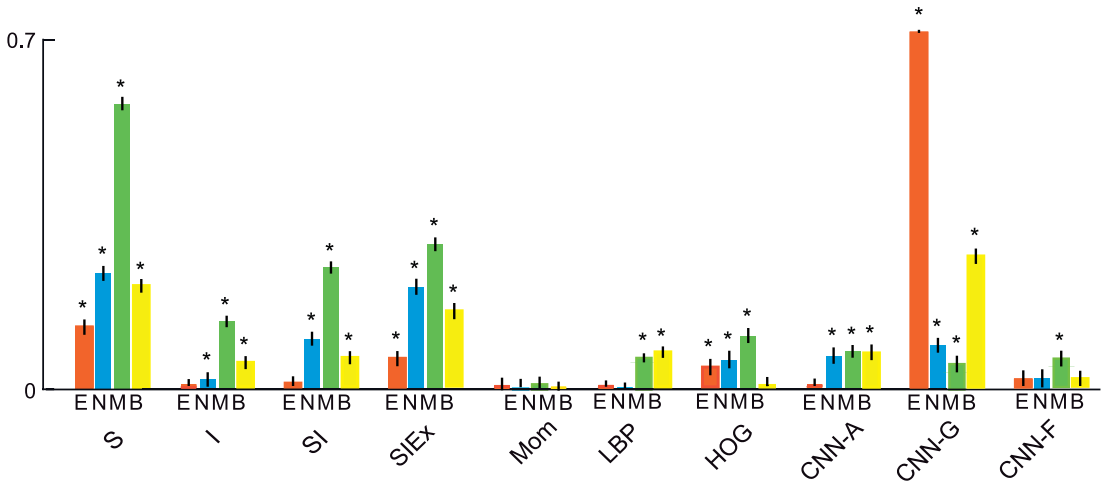

Figure 5. Whole-face model predictions correlate better with mouth shape. The bar plot shows the significant correlations ($p < 0.05$) between predicted geographical-origin labels for models trained on whole-face information with predicted labels for models trained on eyes (E), nose (N), mouth (M), or eyebrow shape (B). The sole exception is CNN-G, which is particularly dependent on eye shape (see text). Predictions based on intensity moment (Mom) did not have significant correlations with those from models trained with eyes, nose, mouth, or eyebrow information. We deemed a correlation significant if less than 5% out of 100 cross-validated splits were below zero.

			Gender (n = 1,647)		Age (n = 459)		Height (n = 218)		Weight (n = 253)	
Feature	Dims	*df*	Accuracy	Rank	Correlation	Rank	Correlation	Rank	Correlation	Rank
S	23	11	0.68 ± 0.00*	12	0.28 ± 0.01*	3	0.32 ± 0.01*	10	0.34 ± 0.01*	8
I	31	15	0.77 ± 0.00*	8	0.21 ± 0.02*	10	0.58 ± 0.02*	4	0.11 ± 0.07*	14
SI	54	24	0.79 ± 0.00*	6	0.26 ± 0.01*	4	0.59 ± 0.02*	3	0.31 ± 0.02*	10
SIex	126	57	0.82 ± 0.00*	4	0.06 ± 0.03*	15	0.36 ± 0.02*	6	0.40 ± 0.02*	3
Mom	7	2	0.58 ± 0.00*	16	0.02 ± 0.02*	16	0.33 ± 0.01*	9	0.03 ± 0.04*	16
LBP	1,328	157	0.55 ± 0.00*	17	**0.36 ± 0.01**	**2**	**0.77 ± 0.01**	**1**	0.37 ± 0.03*	6
HOG	77	**51**	**0.94 ± 0.00**	**1**	**0.38 ± 0.01**	**1**	0.71 ± 0.01*	2	**0.45 ± 0.02**	**2**
CNN-A	512	121	0.79 ± 0.00*	7	0.22 ± 0.03*	7	0.02 ± 0.04*	17	0.25 ± 0.05*	13
CNN-G	512	50	0.80 ± 0.00*	5	0.18 ± 0.03*	11	0.08 ± 0.01*	16	0.29 ± 0.05*	11
CNN-F	4,096	722	0.87 ± 0.01*	2	0.10 ± 0.02*	13	0.10 ± 0.02*	15	0.05 ± 0.12*	15
E	72	5	0.68 ± 0.00*	11	0.00 ± 0.03*	17	0.24 ± 0.02*	13	0.38 ± 0.01*	5
N	66	7	0.68 ± 0.00*	13	0.24 ± 0.02*	5	0.32 ± 0.02*	11	0.32 ± 0.01*	9
M	153	6	0.60 ± 0.00*	15	0.07 ± 0.02*	14	0.25 ± 0.02*	12	0.00 ± 0.07*	17
C	105	5	0.64 ± 0.00*	14	0.23 ± 0.01*	6	0.35 ± 0.01*	7	0.37 ± 0.01*	7
Eb	30	5	0.83 ± 0.00*	3	0.14 ± 0.02*	12	0.24 ± 0.02*	14	0.29 ± 0.01*	12
IP	21	6	0.71 ± 0.00*	9	0.22 ± 0.01*	8	0.36 ± 0.02*	5	**0.47 ± 0.01**	**1**
ENMC	396	16	0.70 ± 0.00*	10	0.21 ± 0.02*	9	0.34 ± 0.02*	8	0.40 ± 0.01*	4

Table 3. Model performance on gender, age, height, and weight prediction. To classify gender, models were trained on the face features together with gender labels. Model accuracy reported is based on 10-fold cross-validation, as before. To predict age, height, and weight, we projected the face features for the available faces into their principal components to account for 95% of the variance, and then performed regularized regression of the features against each attribute. *Asterisks indicate that the given model's performance was below the best model (in bold) on more than 95 of 100 cross-validated splits, which we considered statistically significant. *Notes*: Dims = total number of features; *df* = number of principal-component projections selected for classification/regression; S = spatial features; I = intensity features; SI = spatial and intensity features; SIex = exhaustive spatial and intensity features; Mom = moments of face-pixel intensity; LBP = local binary patterns; HOG = histogram of oriented gradients; CNN-A, CNN-G, CNN-F = deep networks; E = eye; N = nose; M = mouth; C = contour; Eb = eyebrows; IP = interpart distances; ENMC = eye, nose, mouth, and contour together.

shape. The results are summarized in Figure 5. Across models trained on a variety of features, classification performance was best correlated with models based on mouth features. The interesting exception is that the CNN trained for gender classification (CNN-G) is best correlated with eye shape. This particular correlation has been noted before (Binder, Bach, Montavon, Müller, & Samek, 2016), where it was attributed to biases in the training data set and a lack of real-world priors in CNNs trained from scratch with face databases and gender labels.

Predicting gender, age, weight, and height attributes

Humans are adept at judging not only geographical origin but also several other attributes from a given face, such as gender, age, height, and weight. We surmised that there is a common feature representation that can be flexibly reweighted to learn decision boundaries for different attributes. We therefore tested the ability of computational models to predict these attributes. The results are summarized in Table 3.

It can be seen that all models perform gender classification much better than fine-grained geographical-origin classification: The best model for gender was the HOG model, and eyebrow shape (Eb) was the best single part compared to eyes, nose, mouth, contour, and interpart distances.

To a smaller degree, we were also able to predict age and weight using computational models: The best model for age was LBP, and the best single parts were nose (N) and contour (C). The best model for height was also LBP, and the best single parts were interpart distances (IP). The best model for weight was interpart distances (IP). Since we did not collect human estimates for these variables, it is difficult to say whether humans would perform better or worse than these predictions.

Experiment 2: Face-part occlusion

The results of Table 2 show that, among individual face parts such as eyes, nose, and mouth, classifiers trained on mouth features are the most accurate at fine-grained geographical-origin predictions, and

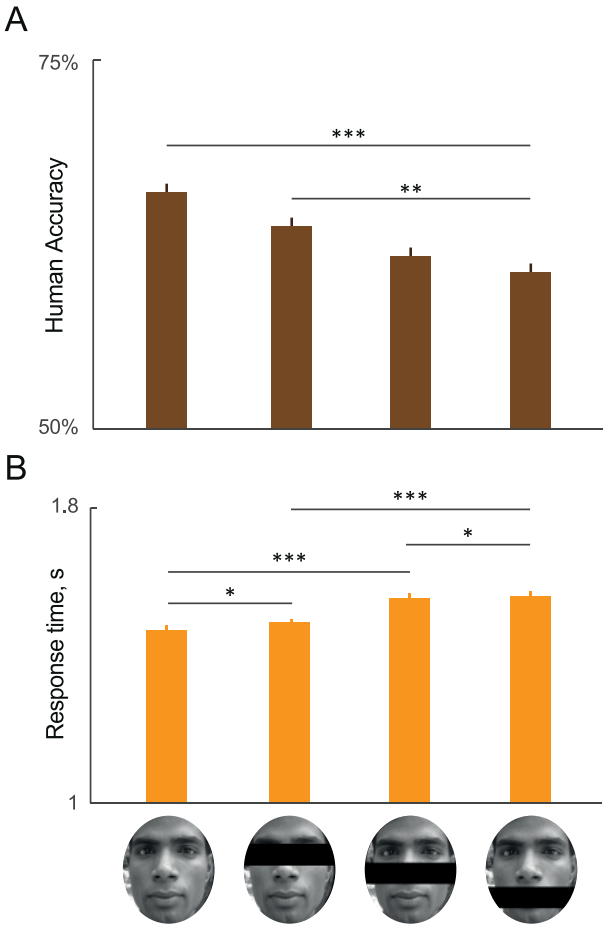

Figure 6. Effect of occluding face parts on classification: (A) Human classification accuracy and (B) response times in each of the occlusion conditions. Error bars indicate standard deviation about the means. Asterisks represent statistical significance: *$p < 0.05$, **$p < 0.005$, and ***$p < 0.0005$ on a rank-sum test performed on binary response-correct labels for all faces concatenated across subjects.

their performance correlates best with human performance.

This in turn predicts that humans base their classification judgements on mouth shape more than on eye or nose shape. We set out to test this prediction using a behavioral experiment on humans. We note that this was by no means a foregone conclusion simply because it was observed using computational analyses: For instance, humans might adaptively use the visible features for classification, thereby maintaining the same accuracy even when a face part is occluded. It could also be that humans use some other complex features based on other face parts that only correlate with a particular face part but can still be extracted when that part is occluded. Testing this prediction in an actual experiment is therefore critical.

Methods

In this experiment, human subjects were asked to perform fine-grained geographical-origin classification on faces in which the eyes, nose, or mouth were occluded in separate blocks. They also performed the same fine-grained classification on unoccluded faces. Importantly, some faces were unique to each block and others were common across blocks. This approach allowed us to compare accuracy for both unique and repeatedly viewed faces across occlusion conditions.
Data set: From the 1,647 faces in the data set, we chose 544 faces spanning moderate (50%) to easy (100%) levels of difficulty, with half the faces being North Indian and the other half South Indian. We then created three occlusion conditions to evaluate the relative importance of the eye, the lower half of the nose, and mouth shape in fine-grained geographical-origin discrimination. Example faces are shown in Figure 6. Our approach is similar in spirit to computational (Zhao, Chellappa, & Rosenfeld, 2003) and behavioral studies (Ellis, Shepherd, & Davies, 1979) that selectively expose face-part information. The occluding band was of the same height in all three cases, and we took care to avoid occluding other face parts (e.g., eyebrows while occluding eyes, or nose while occluding mouth). We then created four sets of faces corresponding to no-occlusion, eye-occluded, nose-occluded, and mouth-occluded conditions. There were a total of 217 faces in each of these conditions, of which 108 were common to all four conditions and 109 were unique to that condition. We ensured during selection that the average human accuracy on North vs. South categorization on the intact versions of each set of 217 faces was comparable and around 69%, based on evaluating accuracy across the full data set.
Subjects: We recruited 24 Indian volunteers (nine women, 15 men; aged 25.7 ± 4.46 years) and obtained informed consent as before. We instructed subjects to indicate geographical-origin labels using key-press responses (N for north and S for south). Each subject was presented with a unique permutation of the $4! = 24$ possible permutation orders of the four occlusion blocks. Only one response was collected for each of the 217 faces shown within a condition.

Results

We analyzed the accuracy of subjects in each occlusion condition separately for the 108 faces common across conditions, as well as for the 109 faces unique to each condition. These data are shown in Figure 6. Subjects were the most accurate on unoccluded faces, as expected (average accuracy = 65.8%). Importantly, classification performance was maximally

impaired for the mouth-occluded (59.8%; $p < 0.0005$ compared to accuracy on unoccluded and $p < 0.005$ compared to accuracy on eye-occluded faces; Wilcoxon rank-sum test performed on binary response-correct labels for all faces concatenated across subjects) and nose-occluded (61.1%; $p = 0.335$ compared to accuracy on mouth-occluded faces; Wilcoxon rank-sum test) conditions, but not as much in the eye-occluded condition (63.6%). Subjects also faced greater difficulty in categorizing mouth-occluded faces, as indicated by longer response times when compared to unoccluded, eye-occluded, and nose-occluded conditions ($p < 0.0005$ compared to unoccluded, $p < 0.05$ compared to eye-occluded or nose-occluded; Wilcoxon rank-sum test performed on response times for all faces concatenated across subjects). We conclude that fine-grained geographical-origin classification depends on mouth shape more than eye or nose shape, thereby validating the prediction from computational modeling.

Experiment 3: Speeded and inverted faces

The results of Experiments 1 and 2 show that mouth shape plays an important role in fine-grained geographical-origin classification in upright faces. In these experiments, upright faces were shown to subjects until they made a response, thereby allowing them to make eye movements to potentially sample face parts. We therefore wondered whether subjects would make qualitatively different responses if they were only shown a briefly flashed face to preclude detailed scrutiny. Likewise, we wondered whether subjects would make qualitatively different responses when they viewed inverted faces, where face features appear in an unfamiliar orientation and configuration. To address these questions, we performed an additional experiment involving upright faces presented for either short or long durations and involving inverted faces.

Methods

Subjects categorized faces presented either in upright, inverted, or speeded conditions. Faces were presented in blocks comprising upright faces shown for 5 s, as in Experiments 1 and 2; upright faces presented for 100 ms followed by a noise mask; and inverted faces shown for 5 s as before. In all three cases the trial continued until a response or 5-s time-out. Subjects were instructed to be fast and accurate, and the order of blocks was counterbalanced across subjects.

Data set: From the 1,647 faces in the data set, we chose 592 faces (296 North, 296 South), with an equal number of male and female faces. This chosen subset of faces was as difficult as the full set of 1,647 faces (median accuracy: 66% for the 592 faces, 58% for all faces; $p = 0.82$; rank-sum test for equal medians).

Subjects: We recruited 25 Indian volunteers (12 women, 13 men; aged 28 ± 5.7 years) and obtained informed consent as before. We instructed subjects to indicate geographical-origin labels using key-press responses (N for north and S for south). Only one response was collected for each unique face in a block. We omitted the data of one participant who later reported not having followed instructions correctly and was at chance in all three blocks

Results

We presented blocks of 592 faces in one of three conditions: upright faces presented until response or for 5 s (regular upright), upright faces presented briefly for 100 ms followed by a noise mask (speeded upright), or inverted faces presented until response or for 5 s. Subjects' performance was above chance in all conditions (average accuracies: 59% for regular upright faces, 57% for speeded upright faces, 55% for inverted faces; $p < 0.05$ for regular and speeded upright, $p < 0.0001$ for inverted faces; chi-square test comparing total correct and wrong responses with a 50/50 split).

Subjects were slightly less accurate in this experiment (average accuracy: 59%) compared to Experiment 1 (average accuracy: 65% on the same faces), presumably because of having to switch blocks. Importantly, they were less accurate on speeded upright faces (average accuracy: 57%) and inverted faces (average accuracy: 55%) compared to regular upright faces (average accuracy: 59%). Subjects were more accurate on regular upright faces than speeded upright faces and inverted faces. These comparisons were statistically significant ($p < 0.05$ in both cases; sign-rank test comparing regular upright with speeded upright or with inverted).

To investigate whether these differences in accuracy were due to speed–accuracy tradeoffs, we compared response times in these three blocks as well. Response times varied systematically across blocks (average: 1.15, 0.95, and 1.31 s, respectively, for regular, speeded, and inverted faces). While subjects were significantly faster on speeded than regular upright faces ($p < 0.0001$; rank-sum test), they were significantly slower on inverted than regular upright faces ($p < 0.0001$; rank-sum test). Thus, the lower accuracy and faster responses to speeded faces compared to regular upright faces stem from a speed–accuracy tradeoff.

Next, we examined the consistency of subjects' responses in the three blocks. To this end we measured the correlation between the average accuracy across faces calculated separately for odd- and even-numbered

			Regular upright		Speeded upright		Inverted	
Human accuracy	-	-	59%	-	57%	-	55%	-
r_{data}	-	-	0.80 ($p < 0.01$)	-	0.83 ($p < 0.01$)	-	0.64 ($p = 0.06$)	-
Best r_{model}	-	-	0.5	-	0.54	-	0.75	-
Feature	Dims	*n*PC	Correlation	Rank	Correlation	Rank	Correlation	Rank
S	23	10	0.28 ± 0.01*	12	0.25 ± 0.01*	12	0.39 ± 0.01*	10
I	31	**15**	**0.40 ± 0.02**	**1**	0.44 ± 0.01	**2**	**0.47 ± 0.01**	**2**
SI	54	**24**	**0.39 ± 0.02**	**2**	0.45 ± 0.02	**1**	**0.47 ± 0.01**	**3**
SIex	126	56	0.29 ± 0.01*	10	0.30 ± 0.01*	8	0.37 ± 0.01*	13
Mom	7	2	0.29 ± 0.01*	8	0.29 ± 0.00*	9	0.41 ± 0.00*	6
LBP	1328	226	0.31 ± 0.02*	6	0.31 ± 0.02*	7	0.41 ± 0.01*	7
HOG	6723	**342**	**0.37 ± 0.02**	**3**	0.39 ± 0.01*	3	0.44 ± 0.02*	5
CNN-A	512	142	0.34 ± 0.01*	4	0.34 ± 0.01*	6	0.46 ± 0.01*	4
CNN-G	512	59	0.30 ± 0.01*	7	0.37 ± 0.01*	5	0.41 ± 0.01*	8
CNN-F	4096	261	0.34 ± 0.01*	5	0.38 ± 0.02*	4	**0.48 ± 0.01**	**1**
E	72	5	0.27 ± 0.01*	14	0.22 ± 0.01*	17	0.37 ± 0.00*	15
N	66	7	0.26 ± 0.01*	17	0.23 ± 0.01*	14	0.36 ± 0.01*	17
M	153	6	0.29 ± 0.01*	9	0.22 ± 0.01*	15	0.37 ± 0.00*	14
C	105	6	0.27 ± 0.01*	13	0.26 ± 0.01*	10	0.40 ± 0.00*	9
Eb	30	4	0.27 ± 0.00*	15	0.22 ± 0.01*	16	0.36 ± 0.00*	16
IP	21	6	0.26 ± 0.01*	16	0.24 ± 0.01*	13	0.38 ± 0.01*	12
ENMC	396	16	0.28 ± 0.01*	11	0.26 ± 0.01*	11	0.38 ± 0.01*	11

Table 4. Model performance on predicting human accuracy in upright, speeded, and inverted faces. To predict human accuracy on upright, speeded, and inverted faces, we projected model features for all 592 faces into their principal components to account for 95% of the variance, and then performed regularized regression. Regular upright and inverted faces were shown for up to 5 s; speeded upright faces were shown for 100 ms followed by a noise mask. *Asterisks indicate that a given model's performance was below the best model (in bold) on more than 95 of 100 cross-validated splits, which we considered statistically significant. *Notes*: Human-pc % = classification accuracy of humans; r_{data} = internal consistency of human responses measured using split-half correlation; Best r_{model} = normalized best model correlation calculated as the ratio of best model correlation and internal consistency of human responses (r_{data}); Dims = total number of features in the model; *n*PC = number of principal components along which features are projected for subsequent regression; S = spatial features; I = intensity features; SI = spatial and intensity features; SIex = exhaustive spatial and intensity features; Mom = moments of face-pixel intensity; LBP = local binary patterns; HOG = histogram of oriented gradients; CNN-A, CNN-G, CNN-F = deep networks; E = eye; N = nose; M = mouth; C = contour; Eb = eyebrows; IP = interpart distances; ENMC = eye, nose, mouth, and contour together.

subjects. This revealed a significant split-half correlation for regular upright faces ($r = 0.80$, $p < 0.01$) and speeded upright faces ($r = 0.83$, $p < 0.01$), and a correlation approaching significance for inverted faces ($r = 0.64$, $p = 0.06$). The split-half correlation for inverted faces was significantly different from both regular and speeded upright faces ($p < 0.0001$ for both comparisons; Fisher's z test).

Next, we asked whether variations in human accuracy across faces could be predicted using computational models as before. The results are summarized in Table 4. As before, models based on spatial and intensity features (SI and I) predicted human accuracy the best across upright, speeded, and inverted faces. Among face parts, mouth shape was the most informative part (compared to eye, nose, and contour) for regular upright faces only, whereas contour was the most informative part for speeded upright and inverted faces.

We also observed an interesting pattern in the ability of the best model to explain human accuracy variations. The correlation between predicted and observed human accuracy for the best model was highest for inverted faces compared to upright faces ($r = 0.40$, 0.45, and 0.47, respectively, for regular upright, speeded upright, and inverted faces). The higher performance of the best model cannot be explained by a difference in response consistency in the three conditions, because the response consistency is in fact higher for upright compared to inverted faces.

To quantify model performance relative to response consistency, we calculated for each set of faces a composite measure of model fit by dividing the best model correlation by the consistency of human subjects on each set. The resulting normalized correlation will have an upper bound of 1 if model predictions are as consistent as humans are with each other. Interestingly, this normalized correlation was larger for inverted faces

compared to upright faces ($r_{norm} = 0.5$, 0.54, and 0.75, respectively, for regular upright, speeded upright, and inverted faces). To assess the statistical significance of these differences, we computed the best model correlation in each of the three conditions using matched fivefold splits and computed the normalized correlation. We repeated this process 1,000 times and counted the number of times the normalized correlation in the inverted condition was smaller than in both upright conditions. This revealed a small fraction of splits with this effect, indicative of statistical significance (fraction $p < 0.001$ for both comparisons).

We conclude that humans rely on more generic face features while categorizing inverted faces compared to upright faces.

General discussion

Here we set out to elucidate the features underlying face perception in humans using a challenging face-classification task involving geographical-origin labels on Indian faces. Our main findings are the following: (a) humans show highly systematic variations in accuracy across faces, indicating that they learn similar feature representations despite widely differing face experience; (b) many computational models achieved human levels of performance, but their error patterns were qualitatively different; (c) a variety of other secondary attributes such as age, gender, and even height and weight were also predictable using computational modeling; (d) mouth shape was the strongest contributor to fine-grained geographical-origin classification compared to other face parts, as evidenced by high accuracy of classifiers trained on face parts; (e) we confirmed this empirically by showing that humans showed the largest drop in accuracy when the mouth was occluded compared to other parts; (f) human performance on inverted-face classification was predicted better by computational models compared to upright faces, suggesting that humans use more generic representations for inverted compared to upright faces. We review each of these findings in the context of the existing literature.

Our main finding is that computational models achieve human levels of performance but show qualitatively different error patterns. This is an important observation for several reasons. It suggests that humans use qualitatively different features. For instance, humans might use features that we did not reliably extract from faces, such as their three-dimensional shape or skin texture. It is also possible that the computer-vision classifiers learn very differently from humans or are susceptible to outliers or noise in face labels. Distinguishing between these possibilities will require systematically training and testing humans on challenging face-classification tasks.

We found mouth shape to be the strongest predictor of fine-grained face classification of North versus South Indian, compared to other face parts. We further confirmed that mouth shape was critical to this classification by comparing human performance on faces with the eyes, nose, or mouth occluded. It was absolutely critical to establish this experimentally: The finding that human classification is best predicted by mouth shape does not guarantee that humans use mouth shape to classify faces. For instance, humans might be using some other feature correlated with mouth shape (but not quantified in this study), in which case occluding the mouth would have had no effect. Likewise, humans may adaptively use whichever features are visible to classify faces, in which case occluding the mouth would leave classification unaffected even though mouth shape is used for unoccluded face classification.

That mouth shape contributes the most to fine-grained geographical-origin classification is consistent with the sensitivity to part shape observed in behavioral readouts of face processing (Valentine, 1991; Abudarham & Yovel, 2016). We have found that face-contour features best predict gender classification, suggesting that different face parts may be informative for different types of judgements. However, these findings do not rule out the contribution of configural features that can interact with local part shape in upright faces (Sergent, 1984; Tanaka & Farah, 1993) or when top and bottom halves of different faces can give rise to new identities in the composite face illusion (Rossion, 2013). On inversion, such interactions can result in new face identities when top and bottom halves of different faces are combined (Young, Hellawell, & Hay, 1987).

It is worth noting that classifiers trained on simpler feature banks yielded comparable or even better performance compared to deep neural networks in our study (Tables 2, 3). This is interesting because a variety of studies have demonstrated a coarse match between object representations across layers in CNNs with those in the visual cortical hierarchy (Yamins et al., 2014), and CNNs have been applied to predict various face attributes such as identity (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014), gender, and age (Levi & Hassner, 2015) with remarkable success. The poor performance of CNNs in predicting human perception may indicate qualitative differences between the two representations despite the overall coarse similarity. This finding also raises the intriguing possibility that face classification in humans may be based on simpler, more interpretable features in contrast to the uninterpretable features used by CNNs (Lapuschkin, Binder, Montavon, Müller, & Samek, 2016; Lapuschkin & Binder, 2017). This finding also suggests that

human accuracy could be used as complementary information to augment and improve face-recognition algorithms (Scheirer et al., 2014), as we have demonstrated for object classification (Katti, Peelen, & Arun, 2019).

Our finding that a variety of other secondary attributes were predictable by computational models demonstrates that the face carries many other interesting signals. While it is not surprising that age, gender, and weight are predictable from the face, it is somewhat more interesting that a whole-body attribute such as height can be predicted by computational models. This finding suggests that there are hidden correlations between the face and the rest of the body. Indeed, face shape is correlated with hand shape (Fink et al., 2005), and a number of other such interesting correlations are emerging from genetic studies (Shaffer et al., 2016; Claes et al., 2018). Face shape may also be modulated by selection pressures that favor diversity in features so as to enable individual identification (Sheehan & Nachman, 2014).

Face inversion is a popular manipulation because it preserves all local image features while altering their orientation and arrangement, and is the basis for striking illusions (Thompson, 2009). We have found that computational models are able to predict human performance on inverted faces better than upright faces. Thus, humans rely on features that are more congruent with those used by computational models when they classify inverted faces. Interestingly, human performance on classifying inverted faces was predicted better by the face contour, whereas upright-face classification was predicted best by the mouth. Taken together, these findings show that humans rely on qualitatively different features for classifying upright and inverted faces.

Finally, our findings are based entirely on faces in front view. Whether the visual system extracts view-specific features or view-invariant features is an important open question (Freiwald & Tsao, 2010; Murty & Arun, 2018). It is possible that humans use entirely different features at different views or prioritize features differently when relevant features are obscured with changing viewpoint. Evaluating these possibilities will require careful testing of face classification with faces shown at varying viewpoints. It is also possible that humans prefer to view faces at specific viewpoints where all features are visible. While this has been observed for objects, whether it is true for faces remains to be tested.

In sum, our results show that humans show highly systematic variations in classification performance on a challenging face-recognition task, and that these variations are qualitatively different from computational models. Our results suggest that discriminative face parts for specific classification tasks can be identified computationally and evaluated experimentally by face-part occlusion.

Keywords: face categorization, ethnicity, computational models, deep convolutional networks, feature representation

Acknowledgments

We thank Pravesh Parekh, Chintan Kaur, N.C. Puneeth, and Arpita Chand for assistance with collecting face images and human behavioral data. We are grateful to Pramod RT and other lab members for help with curating the data set. This work was supported by a DST CSRI postdoctoral fellowship (HK) from the Department of Science and Technology, Government of India, and by Intermediate and Senior Fellowships (500027/Z/09/Z and IA/S/17/1/503081) from the Wellcome Trust–DBT India Alliance (SPA).

Commercial relationships: none.
Corresponding author: Harish Katti.
Email: harish2006@gmail.com.
Address: Centre for Neuroscience, Indian Institute of Science, Bangalore, India.

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3):40, 1–18, https://doi.org/10.1167/16.3.40. [PubMed] [Article]

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 2037–2041, https://doi.org/10.1109/TPAMI.2006.244.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, *23*(2), 193–212, https://doi.org/10.1214/aoms/1177729437.

Barnouti, N. H., Al-Dabbagh, S. S. M., & Matti, W. E. (2016). Face recognition: A literature review. *International Journal of Applied Information Systems*, *11*(4), 21–31, https://doi.org/10.5120/ijais2016451597.

Binder, A., Bach, S., Montavon, G., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. *Lecture*

Notes in Electrical Engineering, *376*, 913–922, https://doi.org/10.1007/978-981-10-0557-2_87.

Brooks, K. R., & Gwinn, O. S. (2010). No role for lightness in the perception of black and white? Simultaneous contrast affects perceived skin tone, but not perceived race. *Perception*, *39*(8), 1142–1145, https://doi.org/10.1068/p6703.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322, https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In M. Valstar, A. French, & T. Pridmore (Eds.), *Proceedings of the British Machine Vision Conference* (pp. 1–12). BMVA Press. https://doi.org/10.5244/C.28.6.

Claes, P., Roosenboom, J., White, J. D., Swigut, T., Sero, D., Li, J., ... Weinberg, S. M. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics*, *50*(3), 414–423, https://doi.org/10.1038/s41588-018-0057-4.

Delaunay, B. (1934). Sur la sphère vide. *Bulletin de l'Académie des Sciences de l'URSS, Classe des Sciences Mathématiques et Naturelles*, *6*, 793–800.

Déniz, O., Bueno, G., Salido, J., & De la Torre, F. (2011). Face recognition using Histograms of Oriented Gradients. *Pattern Recognition Letters*, *32*(12), 1598–1603, https://doi.org/10.1016/j.patrec.2011.01.004.

Duan, X., Wang, C., Liu, X., Li, Z., Wu, J., & Zhang, H. (2010). Ethnic features extraction and recognition of human faces. In *2010 2nd International Conference on Advanced Computer Control* (pp. 125–130). New York, NY: IEEE, https://doi.org/10.1109/ICACC.2010.5487194.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431–439, https://doi.org/10.1068/p080431.

Fink, B., Grammer, K., Mitteroecker, P., Gunz, P., Schaefer, K., Bookstein, F. L., & Manning, J. T. (2005). Second to fourth digit ratio and face shape. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1576), 1995–2001, https://doi.org/10.1098/rspb.2005.3179.

Freiwald, W. A., & Tsao, D. Y. (2010, November 5). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851, https://doi.org/10.1126/science.1194908.

Fu, S., He, H., & Hou, Z.-G. (2014). Race classification from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(12), 2483–2509, https://doi.org/10.1109/TPAMI.2014.2321570.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999, November 11). Signal but not noise changes with perceptual learning. *Nature*, *402*(6758), 176–178, https://doi.org/10.1038/46027.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271, https://doi.org/10.1016/S0042-6989(01)00097-9.

Katti, H., Peelen, M. V., & Arun, S. P. (2017). How do targets, nontargets, and scene context influence real-world object detection? *Attention, Perception, and Psychophysics*, *79*(7), 2021–2036, https://doi.org/10.3758/s13414-017-1359-9.

Katti, H., Peelen, M. V., & Arun, S. P. (2019). Machine vision benefits from human contextual expectations. *Scientific Reports*, *9*(1): 2112, https://doi.org/10.1038/s41598-018-38427-0.

Kozunov, V., Nikolaeva, A., & Stroganova, T. A. (2018). Categorization for faces and tools—two classes of objects shaped by different experience—differs in processing timing, brain areas involved, and repetition effects. *Frontiers in Human Neuroscience*, *11*:650, 1–21, https://doi.org/10.3389/fnhum.2017.00650.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In K.-D. Thoben, M. Busse, B. Denkena, & J. Gausemeier (Eds.), *Advances in Neural Information Processing Systems* (pp. 474–483). London, UK: Elsevier. https://doi.org/10.1016/j.protcy.2014.09.007.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In T. Tuytelaars, L. Fei-Fei, & R. Bajcsy (Eds.), *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2912–2920). https://doi.org/10.1109/CVPR.2016.318.

Lapuschkin, S., & Binder, A. (2017). Understanding and comparing deep neural networks for age and gender classification. In R. Chellappa, A. Hoogs, & Z. Zhang (Eds.), *IEEE International Conference on Computer Vision Workshops* (pp. 1629–1638).

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In H. Bischof, D. Forsyth, C. Schmid, & S. Sclaroff (Eds.), *2015 IEEE Conference on Computer Vision*

and Pattern Recognition Workshops (pp. 34–42), https://doi.org/10.1109/CVPRW.2015.7301352.

Milborrow, S., & Nicolls, F. (2014). Active shape models with SIFT descriptors and MARS. In J. Braz (Ed.), *International Conference on Computer Vision Theory and Applications* (pp. 380–387), https://doi.org/10.5220/0004680003800387.

Murty, N. A. R., & Arun, S. P. (2018). Multiplicative mixing of object identity and image attributes in single inferior temporal neurons. *Proceedings of the National Academy of Sciences, USA*, *115*(14), E3276–E3285, https://doi.org/10.1073/pnas.1714287115.

Ojala, T., Pietikäinen, M., & Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition* (pp. 582–585), https://doi.org/10.1109/ICPR.1994.576366.

O'Toole, A. J., & Natu, V. (2013). Computational perspectives on the other-race effect. *Visual Cognition*, *21*, 1121–1137, https://doi.org/10.1080/13506285.2013.803505.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Procedings of the British Machine Vision Conference 2015* (pp. 41.1–41.12), https://doi.org/10.5244/C.29.41.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253, https://doi.org/10.1080/13506285.2013.772929.

Scheirer, W. J., Anthony, S. E., Nakayama, K., & Cox, D. D. (2014). Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1679–1686, https://doi.org/10.1109/TPAMI.2013.2297711.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, *75*, 221–242.

Setty, S., Husain, M., Behan, P., Gudavalli, J., Kandasamy, M., Vaddi, R., … Rajan, B. (2013). Indian Movie Face Database: A benchmark for face recognition under wide variations. In S. Chaudhury (Ed.), *2013 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics* (pp. 1–5). Jodhpur, India: IEEE, https://doi.org/10.1109/NCVPRIPG.2013.6776225.

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., … Weinberg, S. M. (2016). Genome-wide association study reveals multiple loci influencing normal human facial morphology. *PLoS Genetics*, *12*(8), 1–21, https://doi.org/10.1371/journal.pgen.1006149.

Sharma, R., & Patterh, M. S. (2015). Indian Face Age Database: A database for face recognition with age variation. *International Journal of Computer Applications*, *126*(5), 21–28.

Sheehan, M. J., & Nachman, M. W. (2014). Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature Communications*, *5*, 1–10, https://doi.org/10.1038/ncomms5800.

Somanath, G., Rohith, M., & Kambhamettu, C. (2011). VADANA: A dense dataset for facial image analysis. In D. Metaxas, L. Quan, A. Sanfeliu, & L. V. Gool (Eds.), *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2175–2182), https://doi.org/10.1109/ICCVW.2011.6130517.

Spearman, C. (1910). Correlation calculated from faulty Data. *British Journal of Psychology*, *3*(3), 271–295, https://doi.org/10.1111/j.2044-8295.1910.tb00206.x.

Steinmetz, P. N., & DaSilva, F. (2006). Categorizing blurred images. *Journal of Vision*, *6*(6): 275, https://doi.org/10.1167/6.6.275. [Abstract]

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *46*(2), 225–245, https://doi.org/10.1080/14640749308401045.

Tariq, U., Hu, Y., & Huang, T. S. (2009). Gender and ethnicity identification from silhouetted face profiles. In A. W. Sahlan (Ed.), *Proceedings: International Conference on Image Processing* (pp. 2441–2444). Cairo, Egypt: IEEE, https://doi.org/10.1109/ICIP.2009.5414117.

Thompson, P. (2009). Margaret Thatcher: A new illusion. *Perception*, *38*(6), 483–484, https://doi.org/10.1068/p090483.

Tin, H., & Sein, M. (2011). Race identification for face images. *Proceedings of the International Conference in Computer Engineering*, *1*(2), 2–4.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *43*(2), 161–204, https://doi.org/10.1080/14640749108400966.

Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Wang, M., & Deng, W. (2018). Deep face recognition: A survey. *arXiv*,1804.06655v3.

Wang, Y., Liao, H., Feng, Y., Xu, X., & Luo, J. (2016). Do they all look the same? Deciphering Chinese, Japanese and Koreans by fine-grained deep learning. *arXiv*,1610.01854.

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

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face processing. *Perception*, *16*, 747–759, https://doi.org/10.1068/p160747n.

Zhao, W., Chellappa, R., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, *35*(4), 399–458, https://doi.org/10.1145/954339.954342.