Feature-based face representations and image reconstruction from behavioral and neural data

Adrian Nestor* David C. Plautbc Marlene Behrmannbc

* Department of Psychology at Scarborough, University of Toronto, 1265 Military Trail, Toronto, ON, M1C 1A4. †Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213-3890. ‡Center for the Neural Basis of Cognition, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA, 15213-3890

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The reconstruction of images from neural data can provide a unique window into the content of human perceptual representations. While recent efforts have established the viability of this enterprise using functional magnetic resonance imaging (fMRI) patterns, these efforts have relied on a variety of prespecified image features. Here, we take on the twofold task of deriving features directly from empirical data and of using these features for facial image reconstruction. First, we use a version of reverse correlation to derive visual features from fMRI patterns elicited by a large set of homogeneous face exemplars. Then, we combine these features to reconstruct novel face images from the corresponding neural patterns. This novel approach allows us to estimate collections of facial features associated with different cortical areas as well as to achieve significant levels of reconstruction accuracy. Furthermore, we establish the robustness and the utility of this approach by reconstructing face images from patterns of behavioral data. From a theoretical perspective, the current results provide key insights into the nature of high-level visual representations; and, from a practical perspective, these findings make possible a broad range of image-reconstruction applications via a straightforward methodological approach.

Face recognition relies on visual representations sufficiently complex to distinguish even among highly similar individuals despite considerable variation due to expression, lighting, viewpoint etc. A longstanding conceptual framework, termed face space (1-6), suggests that individual faces are represented in terms of their multidimensional deviation from an "average" face, but the precise nature of the dimensions or features that capture these deviations, and the degree to which they preserve visual detail, remain unclear. Thus, the featural basis of face space along with the neural system that instantiates it remain to be fully elucidated. The present investigation aims not only to uncover fundamental aspects of neural representations but also to establish their plausibility and utility through image reconstruction. Concretely, the current study addresses the issues above in the context of two distinct challenges, first, by determining the visual features employed in face identification and, second, by validating these features through their use in facial image reconstruction.

With respect to the first challenge, recent studies have demonstrated distinct sensitivity to local features (e.g., the size of the mouth) as compared to configural features (e.g., the distance between the eyes and the mouth) in human face-selective cortex (7-10). Also, neurophysiological investigations (1, 11) of monkey cortex have found sensitivity to several facial features, particularly in the eye region of the face. However, most investigations consider only a few handpicked features. Thus, a comprehensive, unbiased assessment of face space still remains to be conducted. Furthermore, most studies target shape at the expense of surface features (e.g., skin tone) despite the relevance of the latter for recognition (12, 13).

With respect to the second challenge, recent studies have taken steps toward image reconstruction from fMRI signals in visual cortex, primarily exploiting low-level visual features (14-16). The recent extension of this work to the reconstruction of face images (17) has demonstrated the promise of exploiting category-specific features (e.g., facial features) associated with activation in higher visual cortex. However, the substantial variability across individual faces in this latter study (due to race, age, image background etc.) limits its conclusions with regard to facial identification and the representations underlying it. Moreover, this attempt deployed prespecified image features due to their reconstruction potential rather than as an argument for their biological plausibility.

The current work addresses the challenges above by adopting a broad, unbiased methodological approach. First, we map cortical areas that exhibit separable patterns of activation to different facial identities. We then construct confusability matrices from behavioral and neural data in these areas to determine the general organization of face space. Next, we extract the visual features accounting for this structure by means of a procedure akin to reverse correlation. And, last, we deploy the very same features for the purpose of face reconstruction. Importantly, our approach relies on an extensive but carefully controlled stimulus set ensuring our focus on fine-grained face identification.

The results of our investigation show that: (i) a range of facial properties such as eyebrow salience and skin tone govern face encoding; (ii) the broad organization of behavioral face space reflects that of its neural homologue, and (iii) high-level face representations retain sufficient detail to support reconstructing the visual appearance of different facial identities from either neural or behavioral data.

Significance

The present work establishes a novel approach to the study of visual representations. This approach allows us to estimate the structure of human face space as encoded by high-level visual cortex, to extract image-based facial features from this structure and to use such features for the purpose of facial image reconstruction. The derivation of visual features from empirical data provides an important step in elucidating the nature and the specific content of face representations. Furthermore, the integrative character of this work sheds new light on the existing concept of face space by rendering it instrumental in image reconstruction. Last, the robustness and generality of our reconstruction approach is established by its ability to handle both neuroimaging and psychophysical data.

Reserved for Publication Footnotes
Pattern-based mapping of facial identity. Participants viewed a set of 120 face images (60 identities × 2 expressions), carefully controlled with respect to both high-level and low-level image properties (see SI text and Fig. 4). Each image was presented at least 10 times per participant across 5 fMRI sessions using a slow event-related design (100 ms stimulus cue, followed by 900 ms stimulus presentation and 7 s fixation). Participants performed a one-back identity task across variation in expression (accuracy was confirmed in the bilateral fusiform face area (FFA) (20) in agreement with previous work (7, 21, 22) but not in the early visual cortex (SI)).

In summary, a total of 8 ROIs localized through pattern-based mapping along with the bilateral FFA were found to be likely candidates for hosting representations of facial identity. Accordingly, these regions formed the basis for the investigation of neural representations reported below.

The similarity structure of human face space. To evaluate the structure of the neural data relevant for identity representation, we extracted the discriminability values of all pairs of facial identities for each of 10 ROIs. Specifically, after mapping these ROIs in each participant, we separately stored, for each participant and ROI, all pairwise discrimination values corresponding to 1770 identity pairs (i.e., all possible pairs derived from 60 identities).

Analogue behavioral measurements were collected in a separate experiment in which participants assessed the behavioral confusability of the stimuli. Briefly, pairs of faces with different expressions were presented sequentially for 400 ms and participants were asked to perform a same/different identity task. Participants were tested with all facial identity pairs across 4 behavioral sessions prior to fMRI scanning. The average accuracy in discriminating each identity pair provided the behavioral counterpart of our neural pattern-discrimination data.

Next, metric multidimensional scaling (MDS) was applied to behavioral and neural discriminability vectors averaged across participants. This analysis forms a natural bridge between recent examinations of neural-based similarity matrices in visual perception (23) and traditional investigations of behavioral-based similarity in the study of face space (2, 24). The outcome of this analysis provides us with the locations of each facial identity in a multidimensional face space. Fig. 1a, c illustrates the distribution of facial identities across the first two dimensions for behavioral data and for right anterior FG data— we focus on this ROI both because of the robustness of its mapping and because of its central role in the processing of facial identity (22). The first two MDS dimensions are particularly relevant since, as detailed below, they contribute important information for reconstruction purposes.

An examination of the results suggests an intuitive clustering of faces based on notable traits such as eyebrow salience. To facilitate the interpretation of these dimensions, faces were separately averaged on each side of the origin proportionally to their discriminability.
Steps involved in the reconstruction procedure: (a) we estimate a multidimensional face space associated with a cortical region and we derive classification image (CI) features for each dimension in CIEL*a*b* color space along with an ‘average’ face (for simplicity, only 2 dimensions are displayed above); (b) we project a new face in this space based on its neural similarity with other faces and we recover its coordinates; (c) we combine CI features proportionately with the coordinates of the new face in order to achieve reconstruction. Thus, as long as we can estimate the position of a stimulus in face space we are able to produce an approximation of its visual appearance.

Examples of face stimuli and their reconstructions from behavioral and fMRI data across (a) neutral and (b) happy expressions. Numbers in the top corners of each reconstruction show its average experimentally-based accuracy (green, left corner) along with its image-based accuracy (red, right corner).

Fig. 4. Examples of face stimuli and their reconstructions from behavioral and fMRI data across (a) neutral and (b) happy expressions. Numbers in the top corners of each reconstruction show its average experimentally-based accuracy (green, left corner) along with its image-based accuracy (red, right corner).

Derivation of facial features underlying face space. The organization of face space is arguably determined by visual features relevant for identity encoding and recognition. An inspection of templates per dimension whose direct comparison informs the perceptual properties encoded by that particular dimension (Fig. 1b, d). The comparison of these templates reveals a host of differences, such as eyebrow thickness, facial hair (i.e., stubble), skin tone, nose shape and mouth height.

Further, the analysis of the behavioral data produced results similar to that of the fMRI data. To evaluate this correspondence, we correlated the coefficients of each facial identity across dimensions extracted for the two data types. This analysis confirmed the similarity between the organization of the first dimensions across behavioral and right aFG data (Fig. 1c, inset) – a broader evaluation of this correspondence is targeted by the assessment of image reconstructions below (see also Fig. S3).

In sum, the present findings verify the presence of consistent structure in our data, assess its impact on the correspondence between behavior and neural processing, and account for this structure in terms of major visual characteristics spanning a range of shape and surface properties.
the results above (Fig. 1) suggests that a simple linear method could capture at least some of these features.

For this purpose, we deployed the following procedure. First, for each dimension, we subtracted each corresponding template from its counterpart, thereby obtaining another template akin to a classification image (CI) (25–27) – that is, a linear estimate of the image-based template that best accounts for identity-related scores along a given dimension (Methods). Then, this template was assessed pixel-by-pixel with respect to a randomly-generated distribution of templates (i.e., by permuting the scores associated with facial identities) in order to reveal pixel values lower or higher than chance (two-tailed permutation test; FDR correction across pixels, q<0.05). These analyses were performed separately for each color channel after converting images to CIEL*a*b* (where L*, a* and b* approximate the lightness, red/green and yellow/blue color-opponent channels of human vision). Examples of raw classification images and their analyses are shown in Fig. 2.

Consistent with our observations above regarding face clustering in face space, the CIs revealed extensive differences in surface properties such as eyebrow thickness and skin tone, as well as shape properties such as nose shape and relative mouth height. For instance, wide bright patches over the forehead and cheeks reflect sensitivity to color differences, whereas thinner patches along the length of the nose and the mouth provide information on shape differences. Also, these differences extend beyond lightness to chromatic channels, whether accompanied or not by similar L* differences. Further, behavioral CIs exhibited larger differences than their neural counterparts. However, most of the ROIs appeared to exhibit some sensitivity to image-based properties.

Thus, our methods were successful in using the similarity structure of neural and behavioral data to derive visual features that capture the topography of human face space.

Facial image reconstruction. An especially compelling way to establish the degree of visual detail captured by a putative set of face space features is to determine the extent to which they support identifiable reconstructions of face images from empirical data. Accordingly, we carried out the following procedure (Fig. 3).

First, we systematically left out each facial identity and estimated the similarity space for the remaining 59 identities. This space is characterized by a set of visual features corresponding to each dimension as well as an average face located at the origin of the space. Second, the left-out identity was projected into this space based on its neural or behavioral similarity with the other faces and its coordinates were retrieved for each dimension. Third, significant features were weighted by the corresponding coordinates and linearly combined along with the average face to generate an image reconstruction. This procedure was carried out separately for behavioral data and for each ROI to generate exemplars with both emotional expression and the anterior temporal lobe (ATL) in identification (28–30).

Discussion
How is facial identity represented by the human visual system? To address this question, we undertook a comprehensive investigation that combines multiple, converging methods in the study of visual recognition, as detailed below.

Cortical mapping of facial identity. Growing sophistication in the analysis of neuroimaging data has facilitated the mapping of the neural correlates of face identification. The examination of face-selective cortex has implicated areas of the FG, STS and the anterior temporal lobe (ATL) in identification (28–30).

Recent investigations relying on multivoxel pattern analysis have extended this work by identifying regions responding with separable patterns of activation to different facial identities, whether or not accompanied by face selectivity (7, 18, 21, 22, 31).

In contrast to previous studies, which have explored the neural code associated with a relatively small number of facial identities, the present study examines the neural and psychological representations underlying an extensive, homogeneous set of unfamiliar faces. This constitutes an exciting test of identity mapping based on fine-grained sensitivity to perceptual differences.

Consistent with previous studies, our investigation found above-chance discrimination in multiple FG, IFG and STS regions as well as in the bilateral PFA. However, the ability of a region to support identity discrimination does not necessarily imply...
that it encodes visual face representations. Higher-level semantic information (32) or even a variety of unrelated task/stimulus properties may be responsible for pattern discrimination (33). The latter possibility is a source of concern especially given certain limitations of the fMRI signal in decoding facial identity (34).

The present findings address this concern by demonstrating that at least certain regions localized via pattern-based mapping contain visual information critical for facial identification. Specifically, three regions of the posterior and anterior FG were able to support identifiable reconstructions of face images. Interestingly, the FFA did not support similar results. However, recent work has shown that the FFA is particularly sensitive to templates driving face detection (35) and can even support the visual reconstruction of such templates (26). Thus, the current results agree with the involvement of the FFA primarily in face detection and, only to a lesser extent, in identification (36, 37). Also, the inability of more anterior regions of the IFG and aMTG to support image reconstruction is broadly consistent with their involvement in processing higher-level semantic information (38).

Thus, our results confirm that facial identification relies on an entire network of cortical regions and, importantly, they point to multiple FG regions as responsible for encoding image-based visual information critical for the representation of facial identity.

**Human face space and its visual features.** What facial properties dominate the organization of human face space? To be clear, our investigation does not target the entirety of face space but rather a specific subdomain: young adult Caucasian male faces. Furthermore, we avoid large differences in appearance due to hair, outer contour or aspect ratio, which are obvious, well-known cues to recognition (2, 39). Instead, we reason that understanding the structure of face representation in a carefully restricted domain challenges the system maximally, and is instrumental in understanding recognition at its best. Our expectation, though, is that the principles revealed here generalize to face recognition as a whole.

A combination of MDS and CI analyses allowed us to assess and visualize the basic organization of face space in terms of both shape and surface properties. Our results reveal a host of visual properties across multiple areas of the face and across different color channels. Notably, we find evidence for the role of eyebrow salience, nose shape, mouth size and positioning, as well as for the role of skin tone. Interestingly, these results agree with previous behavioral work – for instance, the critical role of the eyebrows in individuation has been specifically tested and confirmed (40). Also, several of the properties above appear to be reflected in the structure of both behavioral and neural data.

At the same time, we note that our analyses reveal only a handful of significant features relying on low-dimensional spaces while human face space is believed to be high-dimensional (41). Critically, the number of significant dimensions recovered from the data depends on the signal-to-noise (SNR) ratio of the data as well as on the size of the stimulus set. Here, the restricted number of trials (e.g., 10 presentations per stimulus during scanning) imposes direct limitations on the SNR and allows only the estimation of the most robust features and of the dimensions associated with the entire network of cortical regions and, importantly, the dimensionality of face space but they do open up the possibility of its future investigation with the aid of more advanced imaging techniques and designs.

Last, regarding the presence of visual information across multiple color channels, we note that, traditionally, the role of color in face identification has been downplayed (42, 43). However, recent work has pointed to the value of color in face detection (12) and even in identification (13). As previously suggested (44), color may aid identification when the availability of other cues is diminished. More generally, the difficulty of the task, as induced here by the homogeneity of the stimulus set, could lead to the recruitment of relevant color cues. From a representational standpoint, the current findings suggest that color information is included in the structure of face space and, thus, available when needed. However, additional investigations targeting this hypothesis are needed to ascertain its precise scope and its validity.

In sum, the present results establish the relevance of specific properties, their relative contribution to molding the organization of face space and, conversely, our ability to derive them from neural and behavioral data. More generally, we conclude that specific perceptual representations are encoded in high-level visual cortex and that these representations are fundamentally structured by the visual properties described here.

**Facial identity and image reconstruction.** The fundamental idea grounding our reconstruction method is that, as long as the relative position of a facial identity can be located in face space, its visual appearance can be reconstructed from that of other faces embedded in that space. In a way, our method validates the classic concept of face space (6, 45), by making it instrumental in the concrete enterprise of image reconstruction. Also, from a practical perspective, the presumed benefit of this approach is a more efficient reconstruction method relying on empirically derived representations rather than on hypothetical, prespecified features. For instance, our reconstruction procedure involves only a handful of features (see SI text) whose relevance for representing facial identity is ensured by the process of their derivation and selection. The successful reconstruction of face images drawn from a highly homogeneous set provides strong support for this method.

Overall, the application of reconstruction to both behavioral and fMRI data has extensive theoretical and methodological implications. Theoretically, it points to the general correspondence of face space structure across behavioral and neural data; at the same time, it also highlights the variation of face space across different cortical regions, only some of which contain relevant image-based visual information. Methodologically, the generality and the robustness of the current approach allow its extension to other neuroimaging modalities as well as to data gleaned from patient populations (e.g., to examine distortions of visual representations in prosopagnosia or autism). Thus, our reconstruction results not only provide specific information about the nature of face space, but also open the field to a wide range of future investigations into visual representations and their application to image reconstruction.

In conclusion, our work sheds light on the representational basis of face recognition regarding its cortical locus, its underlying features and their detailed visual content. Our findings reveal a range of shape and surface properties dominating the organization of face space; they show how to synthesize these properties into image-based features of facial identity; they establish a general method for utilizing these features in facial image reconstruction and, last, they validate their behavioral and neural plausibility. More generally, this work demonstrates the strengths of a multi-pronged multivariate paradigm that brings together functional mapping, investigations of behavioral and neural similarity space as well as feature derivation and image reconstruction.

**Methods**

*Footline Author PNAS Issue Date Volume Issue Number | 5*

**Stimuli and design.** A total of 120 images of adult Caucasian males (60 identities x 2 expressions) were selected from multiple face databases and further processed to ensure their homogeneity. Eight right-handed Caucasian adults participated in 9 one-hour experimental sessions (4 behavioral and 5 fMRI). During behavioral sessions, participants viewed pairs of faces, presented in succession, and judged whether they represented the same/different individuals. During fMRI scans, participants performed a continuous 1-back version of the same task using a slow event-related design (8 trials). We imaged 27 oblique slices covering the ventral cortex at 3T (2.5 mm voxels, 21 TR). Informed consent was obtained...
according to procedures approved by the Institutional Review Board of Carnegie Mellon University.

**Pattern-based brain mapping.** Multivoxel pattern-based mapping was performed by walking a spherical searchlight voxel-by-voxel across a cortical map (see Fig. S1a for a group mask). At each location, a single average voxel was used per run for every facial identity. To estimate the neural discriminability linear SVM classification was applied across these patterns for each identity pair using leave-one-run-out cross-validation. These, participant-specific maps were constructed by voxelwise averaging of discrimination estimates across identity pairs. For the purpose of group analysis, all maps were brought into Talairach space and statistical effects were computed across participants (two-tailed t-test against chance).

**Similarity structure analyses.** Behavioral estimates of pairwise face (dis)similarity were computed based on the average discrimination accuracy of each participant during behavioral sessions. Homologous neural-based estimates were computed with the aid of pattern classification in different ROIs based on average discrimination sensitivity. For both data types, this procedure yielded a vector of 1770 pairwise discrimination values.

Further, discrimination vectors were encoded as facial dissimilarity matrices. These matrices were then averaged across participants and analyzed by metric MDS. To interpret the perceptual variation encoded by MDS dimensions (Fig 1a, c), individual faces were averaged on each side of the origin proportionally to their dimension-specific coefficients. The resulting templates were assessed using a reverse correlation approach (25). Concretely, each pair of templates thus obtained were subtracted from each other to derive a classification image (CI) summarizing the perceptual differences specific to that dimension. Each CI was next analyzed, pixel-by-pixel, by comparison with a random set of randomly generated Cs (t-test; q<0.05 correction across pixels). This analysis was separately conducted for the L*, a*, and b* components of face images.

---

Image reconstruction method. For every facial identity an independent estimate of face spaces was constructed through MDS using all other identities. Then, the left out identity is projected in this space via Procrustes alignment. Concretely, the MDS solution derived for all 60 identities is mapped onto the first solution providing us with the coordinates of the target face in the original face space. The resulting coordinates are next used to weight the contribution of significant CIs in the reconstruction process – relevant CIs are selected based on the presence of significant pixels in any of these three channels using a permutation test (FDR correction across pixels; q=0.1). The linear combination of significant CIs along with that of an average face is used to approximate the visual appearance of the target face. This method was conducted separately for 10 ROIs and for behavioral data. Last, a single set of neural reconstructions was derived through the linear combination of ROI-specific reconstructions via an L2 regularized regression model and a leave-one-identity-out procedure.

**Neural and behavioral face space correspondence.** The global correspondence between the two types of face space was assessed by bringing ROI-specific spaces into alignment with behavioral space. Goodness of fit was then estimated via a voxel based analysis of squared errors (SSE) between Procrustes-aligned versions of neural space and behavioral space. Fit estimates were compared to chance across systematic differences in MDS-derived space dimensionality (from 2 to 20 dimensions) via permutation tests. Last, fit estimates across ROIs were compared to each other through parametric tests across participants.

**Acknowledgements.** This research was supported by the National Sciences and Engineering Research Council of Canada (to A.H.), by a Connaught New Researcher Award (to A.N.) and by the National Science Foundation (grant BCS0923763 to M.B. and D.P.; grant SMA-1047155 to M.B. from the Temporary Dynamics of Learning Center).