# Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis

**Adrian Nestor[1], David C. Plaut, and Marlene Behrmann**

Department of Psychology, Carnegie Mellon University, and Center for the Neural Basis of Cognition, Pittsburgh, PA 15213

Edited by Charles G. Gross, Princeton University, Princeton, NJ, and approved May 12, 2011 (received for review February 13, 2011)

**Face individuation is one of the most impressive achievements of our visual system, and yet uncovering the neural mechanisms subserving this feat appears to elude traditional approaches to functional brain data analysis. The present study investigates the neural code of facial identity perception with the aim of ascertaining its distributed nature and informational basis. To this end, we use a sequence of multivariate pattern analyses applied to functional magnetic resonance imaging (fMRI) data. First, we combine information-based brain mapping and dynamic discrimination analysis to locate spatiotemporal patterns that support face classification at the individual level. This analysis reveals a network of fusiform and anterior temporal areas that carry information about facial identity and provides evidence that the fusiform face area responds with distinct patterns of activation to different face identities. Second, we assess the information structure of the network using recursive feature elimination. We find that diagnostic information is distributed evenly among anterior regions of the mapped network and that a right anterior region of the fusiform gyrus plays a central role within the information network mediating face individuation. These findings serve to map out and characterize a cortical system responsible for individuation. More generally, in the context of functionally defined networks, they provide an account of distributed processing grounded in information-based architectures.**

The neural basis of face perception is the focus of extensive research as it provides key insights both into the computational architecture of visual recognition (1, 2) and into the functional organization of the brain (3). A central theme of this research emphasizes the distribution of face processing across a network of spatially segregated areas (4–10). However, there remains considerable disagreement about how information is represented and processed within this network to support tasks such as individuation, expression analysis, or high-level semantic processing.

One influential view proposes an architecture that maps different tasks to distinct, unique cortical regions (6) and, as such, draws attention to the specificity of this mapping (11–20). As a case in point, face individuation (e.g., differentiating Steve Jobs from Bill Gates across changes in expression) is commonly mapped onto the fusiform face area (FFA) (6, 21). Although recent studies have questioned this role of the FFA (14, 15), overall they agree with this task-based architecture as they single out other areas supporting individuation.

However, various distributed accounts have also been considered. One such account ascribes facial identity processing to multiple, independent regions. Along these lines, the FFA's sensitivity to individuation has been variedly extended to areas of the inferior occipital gyrus (5), the superior temporal sulcus (12), and the temporal pole (22). An alternative scenario is that identity is encoded by a network of regions rather than by any of its separate components—such a system was recently described for subordinate-level face discrimination (23). Still another distributed account attributes individuation to an extensive ventral cortical area rather than to a network of smaller separate regions (24). Clearly, the degree of distribution of the information supporting face individuation remains to be determined.

Furthermore, insofar as face individuation is mediated by a network, it is important to determine how information is distributed across the system. Some interesting clues come from the fact that right fusiform areas are sensitive to both low-level properties of faces (16, 25) and high-level factors (26, 27), suggesting that these areas may mediate between image-based and conceptual representations. If true, such an organization should be reflected in the pattern of information sharing among different regions.

The current work investigates the nature and the extent of identity-specific neural patterns in the human ventral cortex. We examined functional MRI (fMRI) data acquired during face individuation and assessed the discriminability of activation patterns evoked by different facial identities across variation in expression. To uncover the neural correlate of identity recognition, we performed dynamic multivariate mapping by combining information-based mapping (28) and dynamic discrimination analysis (29). The results revealed a network of fusiform and anterior temporal regions that respond with distinct spatiotemporal patterns to different identities. To elucidate the distribution of information, we examined the distribution of diagnostic information across these regions using recursive feature elimination (RFE) (30) and related the information content of different regions to each other. We found that information is evenly distributed among anterior regions and that a right fusiform region plays a central role within this network.

## Results

Participants performed an individuation task with faces (Fig. 1) and orthographic forms (OFs) (Fig. S1). Specifically, they recognized stimuli at the individual level across image changes introduced by expression (for faces) or font (for OFs). Response accuracy was at ceiling (>95%) as expected given the familiarization with the stimuli before scanning and the slow rate of stimulus presentation. Thus, behavior exhibits the expected invariance to image changes, and the current investigation focuses on the neural codes subserving this invariance.

**Dynamic Multivariate Mapping.** The analysis used a searchlight (SL) with a 5-voxel radius and a 3-TR (Time to Repeat) temporal envelope to constrain spatiotemporal patterns locally. These patterns were submitted to multivariate classification on the basis of facial identity (*Methods* and *SI Text*). The outcome of the analysis is a group information-based map (28) revealing the strength of discrimination (Fig. 2). Each voxel in this map represents an entire region of neighboring voxels defined by the SL mask.

Author contributions: A.N., D.C.P., and M.B. designed research; A.N. performed research; A.N., D.C.P., and M.B. analyzed data; and A.N., D.C.P., and M.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The fMRI dataset has been deposited with the XNAT Central database under the project name "STSL."

[1]To whom correspondence should be addressed. E-mail: anestor@andrew.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1102433108/-/DCSupplemental.

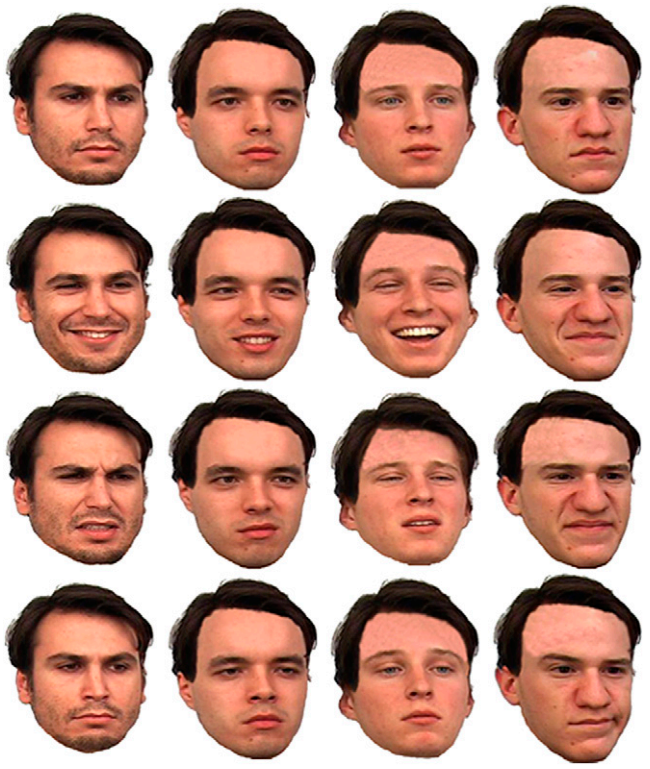

**Fig. 1.** Experimental face stimuli (4 identities × 4 expressions). Stimuli were matched with respect to low-level properties (e.g., mean luminance), external features (hair), and high-level characteristics (e.g., sex). Face images courtesy of the Face-Place Face Database Project (http://www.face-place.org/) Copyright 2008, Michael J. Tarr. Funding provided by NSF Award 0339122.

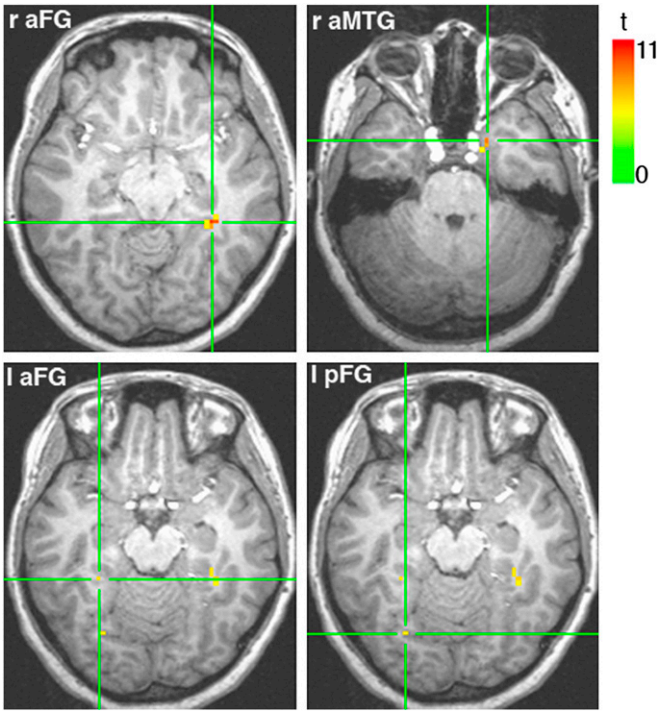


**Fig. 2.** Group information-based map of face individuation. The map is computed using a searchlight (SL) approach and estimates the discriminability of facial identities across expression ($q < 0.05$). Each voxel in the map represents the center of an SL-defined region supporting identity discrimination. The four slices show the sensitivity peaks of the four clusters revealed by this analysis.

Our mapping revealed four areas sensitive to individuation: two located bilaterally in the anterior fusiform gyrus (aFG), one in the right anterior medial temporal gyrus (aMTG), and one in the left posterior fusiform gyrus (pFG). The largest of these areas corresponded to the right (r)aFG whereas the smallest corresponded to its left homolog (Table 1).

To test the robustness of our mapping, the same cortical volume (Fig. S2*A*) was examined with SL masks of different sizes. These alternative explorations produced qualitatively similar results (Fig. S3 *A–C*). In contrast, a univariate version of the mapping (*SI Text*) failed to uncover any significant regions, even at a liberal threshold ($q < 0.10$), attesting to the strength of multivariate; mapping.

To further evaluate these results, we projected the four areas from the group map back into the native space of each subject and expanded each voxel to the entire SL region centered on it. The resulting SL clusters (Fig. S3*D*) mark entire regions able to support above-chance classification. Examination of subject-specific maps revealed that the bilateral aFG clusters were consistently located anterior to the FFA [rFFA peak coordinates, 39, −46, and −16; left (l)FFA, −36, −47, and −18]. However, we also found they consistently overlapped with the FFA (mean ± SD: 24 ± 14% of raFG volume and 35 ± 20% of laFG).

Finally, multivariate mapping was applied to other types of discrimination: expression classification (across identities) and category-level classification (faces versus OFs). Whereas the former analysis did not produce any significant results, the latter found reliable effects extensively throughout the cortical volume analyzed (Fig. S2*B*). In contrast, a univariate version of the latter analysis revealed considerably less sensitivity than its multivariate counterpart (Fig. S2*C*).

The results above suggest that identity coding relies on a distributed cortical system. Clarifying the specificity of this system to face individuation is addressed by our region-of-interest (ROI) analyses.

**ROI Analyses.** First, we examined whether the FFA supports reliable face individuation as tested with pattern classification. Bilateral FFAs were identified in each subject using a standard face localizer, and discrimination was computed across all features in a region—given the use of spatiotemporal patterns, our features are voxel *X* time-point pairings rather than voxels alone. The analysis revealed above-chance performance for rFFA (Fig. 3).

To reduce overfitting, we repeated the analysis above using subsets of diagnostic features identified by multivariate feature selection, specifically RFE. The method works by systematically removing features, one at a time, on the basis of their impact on classification (*SI Text*). Following this procedure, we found above-chance performance bilaterally in the FFA (Fig. 3). In contrast, early visual cortex (EVC) did not exhibit significant sensitivity either before or after feature selection.

Second, we reversed our approach by using multivariate mapping to localize clusters and univariate analysis to assess face se-

**Table 1. Areas sensitive to face individuation**

| | Coordinates (peak) | | | | |
|---|---|---|---|---|---|
| Region | *x* | *y* | *z* | SL centers (voxels) | Peak *t* value |
| raFG | 33 | −39 | −9 | 16 | 11.31 |
| raMTG | 19 | 6 | −26 | 8 | 9.90 |
| lpFG | −26 | −69 | −14 | 2 | 7.11 |
| laFG | −29 | −39 | −14 | 1 | 7.24 |

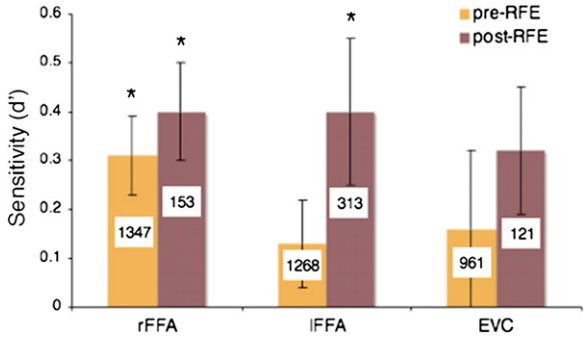


**Fig. 3.** Sensitivity estimates in three ROIs. Facial identity discrimination was computed using both the entire set of features in an ROI and a subset of diagnostic features identified by multivariate feature selection (i.e., RFE), the two types of classification are labeled as pre- and post-RFE. The average number of features involved in classification is superimposed on each bar. The results indicate that the bilateral FFA, in contrast to an early visual area, contains sufficient information to discriminate identities above chance ($P < 0.05$).

lectivity. Concretely, we examined the face selectivity of the SL clusters using the data from our functional localizers. No reliable face selectivity was detected for any cluster.

Third, SL clusters along with the FFA were tested for their ability to discriminate expressions across changes in identity. Above-chance discrimination was found in rFFA and raMTG ($P < 0.05$).

Finally, we tested our clusters for OF individuation across variation in font. The analysis found sensitivity in two regions: rFFA and lpFG.

These findings are important in several respects. They suggest conventionally defined face selectivity, although informative, may not be enough to localize areas involved in fine-level face representation. Also, they show that the identified network is not exclusively dedicated to individuation or even to face processing per se. One hypothesis, examined below, may explain this involvement in multiple types of perceptual discrimination simply by appeal to low-level image properties.

**Impact of Low-Level Image Similarity on Individuation.** To determine the engagement of the network in low-level perceptual processing, image similarity was computed across images of different individuals using an $L_2$ metric (Table S1). For each pair of face identities, the average distance was correlated with the corresponding discrimination score produced by every ROI (including the FFA). The only ROI susceptible to low-level image sensitivity was laFG ($P < 0.05$ uncorrected) (Fig. S4).

These results, along with the inability of the EVC to support individuation, suggest that low-level similarity is unlikely to be the main source of the individuation effects observed here.

**Feature Ranking and Mapping.** Having located a network of regions sensitive to face identity, we set out to determine the spatial and temporal distribution of the features diagnostic of face individuation. Specifically, we performed RFE analysis jointly across all SL clusters and recorded the ranking of the features within each subject. To eliminate any spatial bias within the initial feature set, we started with an equally large number of features for each cluster: the 1,000 top-ranked ones based on single-cluster RFE computation.

Fig. 4*A* shows the average ranking of the 400 most informative features across subjects and Fig. 4*B* summarizes their distribution across clusters and time points. We found a significant effect of cluster (two-way analysis of variance, $P < 0.05$) but no effect of time point and no interaction. Further comparisons revealed that lpFG contains fewer features than other clusters ($P < 0.05$), which did not differ among each other. The time course of feature elimination revealed that lpFG features were consistently eliminated at a higher rate than features from other clusters (Fig. 4*C*). A

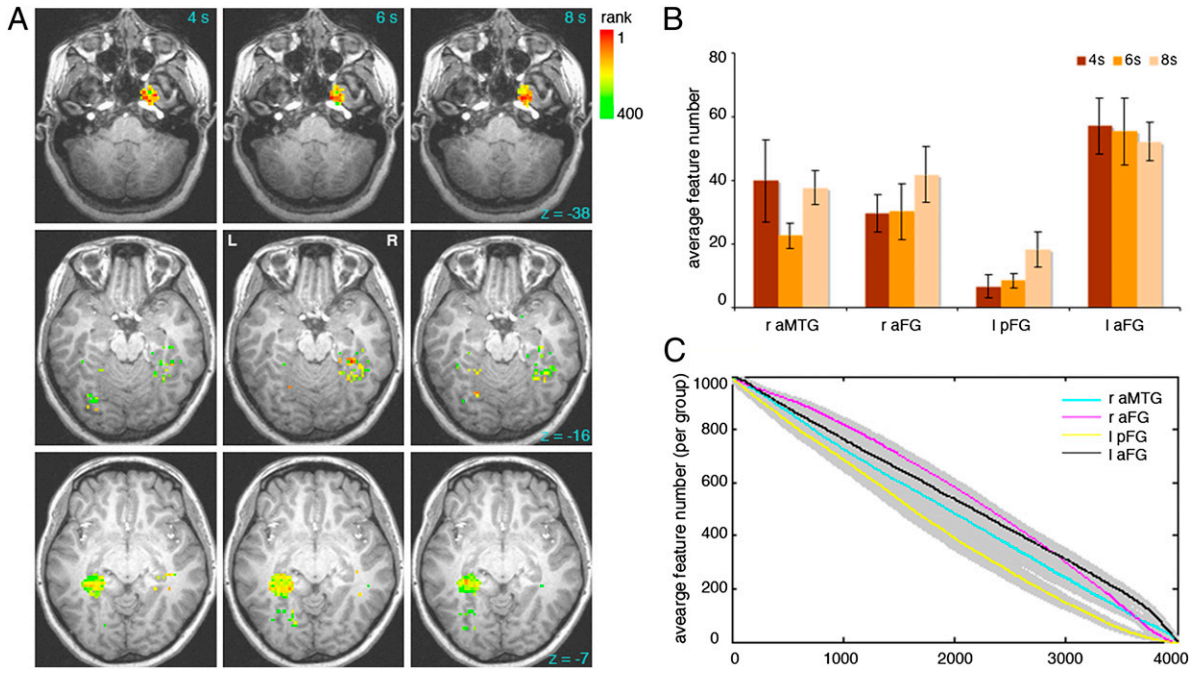


**Fig. 4.** Spatiotemporal distribution of information diagnostic for face individuation. (*A*) Group map of average feature ranking for the top 400 features—rows show different slices and columns different time points. Color codes the ranking of the features across space (the four regions identified by our SL analysis) and time (4–8 s poststimulus onset). The map shows a lower concentration of features in the lpFG relative to other regions but a comparable number of features across time. (*B*) Average feature distribution across subjects by cluster and time point (the bar graph quantifies the results illustrated in *A*). (*C*) Time course of feature elimination by ROI for 4,000 features (top 1,000 features for each ROI). This analysis confirms that lpFG features are eliminated at a higher rate, indicative of their reduced diagnosticity (shaded areas show ±1 SE across subjects).

similar analysis across time points showed no substantial differences across time (Fig. S5).

Feature mapping provides a bird's eye view of information distribution across regions. In our case, it reveals a relatively even division of diagnostic information among anterior regions. Further pairwise comparisons of SL clusters were deployed to examine how areas share information with each other.

**Information-Based Pairwise Cluster Analysis.** Whereas activation patterns in different areas are not directly comparable with each other (e.g., because they have different dimensionalities, there is no obvious mapping between them, etc.), the classification results they produce serve as convenient proxies (*Methods*). Here, we compared classification patterns for pairs of regions while controlling for the pattern of correct labels. Thus, the analysis focuses on common biases in misclassification.

First, we computed the partial correlation between classification patterns corresponding to different clusters. Group results are displayed as a graph in Fig. 5. Similarity scores for all pairs of regions tested above chance (one-sample $t$ test, $P < 0.01$). Similarity scores within the network were not homogeneous (one-way analysis of variance, $P < 0.05$) mainly because raFG evinced higher similarity estimates than the rest ($P < 0.01$). In addition, we computed similarity scores between the four clusters and the EVC. Average similarity scores within the network were significantly higher than scores with the EVC (paired-sample $t$ test, $P < 0.01$). Second, to verify our findings using an information-theoretic measure, we computed the conditional mutual information between classification patterns produced by different clusters (Fig. 5). Examination of information estimates revealed a relational structure qualitatively similar to that obtained using correlation.

We interpret the results above as evidence for the central role of the right FG in face individuation. More generally, they support the idea of redundant information encoding within the network.

## Discussion

The present study investigates the encoding of facial identity in the human ventral cortex. Our investigation follows a multivariate approach that exploits multivoxel information at multiple stages from mapping to feature selection and network analysis. We favor this approach because multivariate methods are more successful at category discrimination than univariate tests (31) and possibly more sensitive to "subvoxel" information than adaptation techniques (32). In addition, we extend our investigation to take advantage of spatiotemporal information (29) and, thus, optimize the discovery of small fine-grained pattern differences underlying the perception of same-category exemplars.

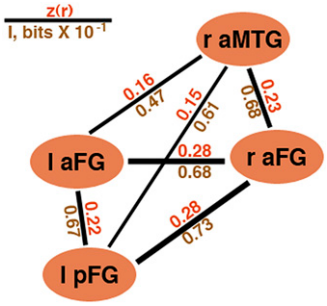


**Fig. 5.** Pairwise ROI relations. The pattern of identity (mis)classifications is separately compared for each pair of regions using correlation-based scores (red) and mutual information (brown). Specifically, we relate classification results across regions while controlling for the pattern of true labels. These measures are used as a proxy for assessing similarity in the encoding of facial identity across regions. Of the four ROIs, the raFG produced the highest scores in its relationship with the other regions (connector width is proportional to $z$ values).

**Multiple Cortical Areas Support Face Individuation.** Multivariate mapping located four clusters in the bilateral FG and the right aMTG encoding facial identity information. These results indicate that individuation relies on a network of ventral regions that exhibit sensitivity to individuation independently of each other. This account should be distinguished both from local ones (6, 21) and from other versions of distributed processing (23, 24).

With regard to the specific clusters identified, previous work uncovered face-selective areas in the vicinity of the FFA, both posterior (9) and anterior (33) to it. Our clusters did not exhibit face selectivity when assessed with a univariate test. However, this lack of selectivity may reflect the variability of these areas (9, 33) and/or the limitations of univariate analysis (34). Alternatively, it is possible that face processing does not necessarily entail face selectivity (35). More relevantly here, face individuation, rather than face selectivity, was previously mapped to an area in the anterior vicinity of the FFA (14). Overall, our present results confirm the involvement of these areas in face processing and establish their role in individuation.

On a related note, the proximity of these fusiform clusters to the FFA may raise questions as to whether they are independent clusters or rather extensions of the FFA (7, 9). On the basis of differences in peak location and the lack of face selectivity, we treat them here as distinct from the FFA although further investigation is needed to fully understand their relationship with this area.

Unlike the FG areas discussed above, an anterior region of the right middle temporal cortex (36, 37) or temporal pole (22) was consistently associated with identity coding. Due to its sensitivity to higher-level factors, such as familiarity (22), and its involvement in conceptual processing (38, 39), the anterior temporal cortex is thought to encode biographical information (6). Whereas our ability to localize this region validates our mapping methodology, the fact that our stimuli were not explicitly associated with any biographical information suggests that the computations hosted by this area also involve a perceptual component. Consistent with this, another mapping attempt, based on perceptual discrimination (15), traced face individuation to a right anterior temporal area. Also, primate research revealed encoding of face–space dimensions in the macaque anterior temporal cortex (40) as well as different sensitivity to perceptual and semantic processing of facial identity (41). In light of these findings, we argue that this area is part of the network for *perceptual* face individuation although the representations it hosts are likely to also comprise a conceptual component.

In sum, our mapping results argue for a distributed account of face individuation that accommodates a multitude of experimental findings. Previous imaging research may have failed to identify this network due to limits in the sensitivity of the methods used in relation with the size of the effect. Inability to find sensitivity in more than one region can easily lead to a local interpretation. At the same time, combining information from multiple regions may counter the limitations of one method but overestimate the extent of distributed processing. Our use of dynamic multivariate mapping builds upon these previous findings and is a direct attempt to increase the sensitivity of these mapping methods at the cost of computational complexity.

Importantly, the regions uncovered by our analysis may represent only a subset of the full network of regions involved in identity processing. We allow for this possibility given our limited coverage (intended to boost imaging resolution in the ventral cortex) as well as the lower sensitivity associated with the imaging of the inferior temporal cortex. In particular, regions of the superior temporal sulcus and prefrontal cortex (4, 10) are plausible additions to the network uncovered here.

**Individuation Effects Are Not Reducible to Low-Level Image Processing.** To distinguish identity representations from low-level image differences, we appealed to a common source of intraindividual image variations: emotional expressions. Also, identity was not predictable in our case by prominent external features, such as hair, or by image properties associated with higher-level characteristics (e.g., sex or age). Furthermore, we assessed the contribution of interindividual low-level similarity to discrimination performance. Of all regions examined, only laFG showed potential reliance on low-level properties. Finally, an examination of an early visual area did not produce any evidence for identity encoding. Taken together, these results render unlikely an explanation of individuation effects based primarily on low-level image properties.

**The Face Individuation System Does Not Support General-Purpose Individuation.** To address the domain specificity of the system identified, we examined whether "abstract" OFs (i.e., independent of font) can be individuated within the regions mapped for faces. Although highly dissimilar from faces, OFs share an important attribute with them, by requiring fine-grained perceptual discrimination at the individual level. In addition, they appear to compete with face representations (42) and to rely on similar visual processing mechanisms (43). Thus, they may represent a more suitable contrast category for faces than other familiar categories, such as houses. Our attempt to classify OF identities revealed that two regions, the lpFG and the rFFA, exhibited sensitivity to this kind of OF information.

Furthermore, to examine the task specificity of the network, we evaluated the ability of its regions to support expression discrimination and found that the rFFA, along with the aMTG, was able to perform this type of discrimination.

Thus, it appears that the mapped network does not support general object individuation, although it may share resources with the processing of other tasks as well as of other visual categories.

**FFA Responds to Different Face Identities with Different Patterns.** The FFA (44, 45) is one of the most intensely studied functional areas of the ventral stream. However, surprisingly, its role in face processing is far from clear. At one extreme, its involvement in face individuation (6, 18, 21) has been called into question (14–16); at the other, it has been extended beyond the face domain to visual expertise (17) and even to general object individuation (13).

Previous studies of the FFA using multivariate analysis have not been successful in discovering identity information (15, 24) or even subordinate-level information (23) about faces. The study of patient populations is also not definitive. FG lesions associated with acquired prosopagnosia (46) adversely impact face individuation, confirming the critical role of the FFA. However, individuals with congenital prosopagnosia appear to exhibit normal FFA activation profiles (47) in the presence of compromised fiber tracts connecting the FG to anterior areas (48). Thus, the question is more pertinent than ever: Does the FFA encode identity information?

Our results provide evidence that the FFA responds consistently across different images of the same individual, but distinctly to different individuals. In addition, we show that the right FFA can individuate OFs and decode emotional expressions, consistent with its role in expression recognition (12).

Thus, we confirm the FFA's sensitivity to face identity using pattern analysis. Moreover, we show that it extends beyond both a specific task, i.e., individuation, and a specific domain, i.e., faces. Further research is needed to determine how far its individuation capabilities extend and how they relate with each other.

**Informative Features Are Evenly Distributed Across Anterior Regions.** How uniformly is information distributed across multiple regions? At one extreme, the system may favor robustness as a strategy and assign information evenly across regions. At the other, its structure may be shaped by the feed-forward flow of information and display a clear hierarchy of regions from the least to the most diagnostic. The latter alternative is consistent, for instance, with a posterior-to-anterior accumulation of information culminating in the recruitment of the aMTG as the endpoint of identity processing.

Our results fall in between these two alternatives. Anterior regions appear to be at an advantage compared with the left pFG. At the same time, there was no clear differentiation among anterior regions in terms of the amount of information represented, suggesting that information is evenly distributed across them.

**The Right aFG May Be a Hub in the Facial Identity Network.** As different regions are not directly comparable as activation patterns, we used their classification results as a proxy for their comparison. Using this approach, we found that different regions do share information with each other, consistent with redundancy in identity encoding. Furthermore, we found that pairwise similarities are more prominent in relation with the right aFG than among other network regions, suggesting that the raFG plays a central role within the face individuation network. Thus, the raFG mirrors the role played by the right FFA among face-selective regions as revealed by functional connectivity (4). One explanation of this role is that a right middle/anterior FG area serves as an interface between low- and high-level information.

Two lines of evidence support this hypothesis. Recent results show, surprisingly, that the FFA exhibits sensitivity to low-level face properties (16, 25). Additionally, the right FG is subject to notable top–down effects (26, 27). Maintaining a robust interface between low-level image properties and high-level factors is likely a key requirement for fast, reliable face processing. Critical for our argument, this requirement would lead to the formation of an FG activation/information hub. Future combinations of information and activation-based connectivity analyses might be able to assess such hypotheses and provide full-fledged accounts of the flow of information in cortical networks.

**Summary.** A broad body of research suggests that face perception relies on an extensive network of cortical areas. Our results show that a single face-processing task, individuation, is supported by a network of cortical regions that share resources with the processing of other visual categories (OFs) as well as other face-related (expression discrimination) tasks. Detailed investigation of this network revealed an information structure dominated by anterior cortical regions, and the right FG in particular, confirming its central role in face processing. Finally, we suggest that a full understanding of the operation of this system requires a combination of conventional connectivity analyses and information-based explorations of network structure.

## Methods

An extended version of this section is available in *SI Text*.

**Design.** Eight subjects were scanned across multiple sessions using a slow event-related design (10-s trials). Subjects were presented with a single face or OF stimulus for 400 ms and were asked to identify the stimulus at the individual level using a pair of response gloves. We imaged 27 oblique slices covering the ventral cortex at 3T (2.5-mm isotropic voxels, 2-s TR).

**Dynamic Information-Based Brain Mapping.** The SL was walked voxel-by-voxel across a subject-specific cortical mask. The mask covered the ventral cortex (Fig. S2) and was temporally centered on 6-s poststimulus onset. At each location within the mask, spatiotemporal patterns (29) were extracted for each stimulus presentation. To boost their signal, these patterns were averaged within runs on the basis of stimulus identity. Pattern classification was performed using linear support vector machines (SVM) with a trainable $c$ term followed by leave-one-run-out cross-validation. Classification was separately applied to each pair of identities (six pairs based on four identities). Discrimination performance for each pair was encoded using $d'$ and an

average information map (28) was computed across all pairs. For the purpose of group analysis, these maps were normalized into Talairach space and examined for above-chance sensitivity ($d' > 0$), using voxelwise $t$ tests across subjects [false discovery rate (FDR) corrected]. Expression discrimination followed a similar procedure.

Analyses were carried out in Matlab with the Parallel Processing Toolbox running on a ROCKS+ multiserver environment.

**ROI Localization.** Three types of ROIs were localized as follows: (*i*) We identified locations of the group information map displaying above-chance discriminability and projected their coordinates in the native space of each subject. ROIs were constructed by placing spherical masks at each of these locations—a set of overlapping masks gave rise to a single ROI. (*ii*) Bilateral FFAs were identified for each subject by standard face–object contrasts. ROIs were constructed by applying a spherical mask centered on the FG face-selective peak. (*iii*) Anatomical masks were manually drawn around the calcarine sulcus of each subject and a mask was placed at the center of these areas. The results serve as a rough approximation of EVC. All masks used for ROI localization had a 5-voxel radius.

**RFE-Based Analysis.** SVM-based RFE (30) was used for feature selection and ranking—the order of feature elimination provides an estimate of feature diagnosticity for a given type of discrimination. To obtain unbiased estimates of performance, we executed two types of cross-validation. We performed cross-validation, first, at each RFE iteration step to measure performance and, second, across iteration steps to find the best number of features. RFE analysis was applied to all types of ROIs described above.

**Pairwise ROI Analysis.** We computed the similarity of classification patterns produced by each pair of ROIs, namely the patterns of classification labels obtained with test instances during cross-validation. However, patterns are likely correlated across regions by virtue of the ability of SVM models to approximate true labels. Therefore, we measured pattern similarity with partial correlation while controlling for the pattern of true labels. Correlations were computed for each pair of facial identities, transformed using Fisher's $z$, and averaged within subjects. Critically, to eliminate common biases based on spatial proximity between regions (because of spatially correlated noise) all activation patterns were $z$-scored before classification. To obtain estimates of the information shared between ROIs we also computed the conditional mutual information between ROI-specific classification patterns given the pattern of true labels.

**ACKNOWLEDGMENTS.** This work was funded by National Science Foundation Grant SBE-0542013 to the Temporal Dynamics of Learning Center and by National Science Foundation Grant BCS0923763. M.B. was partially supported by a Weston Visiting Professorship at the Weizmann Institute of Science.

1. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11: 333–341.
2. Jiang X, et al. (2006) Evaluation of a shape-based model of human face discrimination using FMRI and behavioral techniques. *Neuron* 50:159–172.
3. Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc Natl Acad Sci USA* 107:11163–11170.
4. Fairhall SL, Ishai A (2007) Effective connectivity within the distributed cortical network for face perception. *Cereb Cortex* 17:2400–2406.
5. Gauthier I, et al. (2000) The fusiform "face area" is part of a network that processes faces at the individual level. *J Cogn Neurosci* 12:495–504.
6. Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. *Trends Cogn Sci* 4:223–233.
7. Pinsk MA, et al. (2009) Neural representations of faces and body parts in macaque and human cortex: A comparative FMRI study. *J Neurophysiol* 101:2581–2600.
8. Rossion B, et al. (2003) A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain* 126: 2381–2395.
9. Weiner KS, Grill-Spector K (2010) Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *Neuroimage* 52:1559–1573.
10. Tsao DY, Schweers N, Moeller S, Freiwald WA (2008) Patches of face-selective cortex in the macaque frontal lobe. *Nat Neurosci* 11:877–879.
11. Calder AJ, Young AW (2005) Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci* 6:641–651.
12. Fox CJ, Moon SY, Iaria G, Barton JJ (2009) The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *Neuroimage* 44:569–580.
13. Haist F, Lee K, Stiles J (2010) Individuating faces and common objects produces equal responses in putative face-processing areas in the ventral occipitotemporal cortex. *Front Hum Neurosci* 4:181.
14. Nestor A, Vettel JM, Tarr MJ (2008) Task-specific codes for face recognition: How they shape the neural representation of features for detection and individuation. *PLoS ONE* 3:e3978.
15. Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA* 104: 20600–20605.
16. Xu X, Yue X, Lescroart MD, Biederman I, Kim JG (2009) Adaptation in the fusiform face area (FFA): Image or person? *Vision Res* 49:2800–2807.
17. Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci* 3:191–197.
18. Freiwald WA, Tsao DY, Livingstone MS (2009) A face feature space in the macaque temporal lobe. *Nat Neurosci* 12:1187–1196.
19. Andrews TJ, Ewbank MP (2004) Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage* 23:905–913.
20. Pourtois G, Schwartz S, Seghier ML, Lazeyras F, Vuilleumier P (2005) View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: An event-related fMRI study. *Neuroimage* 24:1214–1224.
21. Kanwisher N, Yovel G (2006) The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci* 361:2109–2128.
22. Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci* 8:107–113.
23. Op de Beeck HP, Brants M, Baeck A, Wagemans J (2010) Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *Neuroimage* 49:3414–3425.
24. Natu VS, et al. (2010) Dissociable neural patterns of facial identity across changes in viewpoint. *J Cogn Neurosci* 22:1570–1582.
25. Yue X, Cassidy BS, Devaney KJ, Holt DJ, Tootell RB (2011) Lower-level stimulus features strongly influence responses in the fusiform face area. *Cereb Cortex* 21:35–47.
26. Bar M, et al. (2006) Top-down facilitation of visual recognition. *Proc Natl Acad Sci USA* 103:449–454.
27. Cox D, Meyers E, Sinha P (2004) Contextually evoked object-specific responses in human visual cortex. *Science* 304:115–117.
28. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
29. Mourão-Miranda J, Friston KJ, Brammer M (2007) Dynamic discrimination analysis: A spatial-temporal SVM. *Neuroimage* 36:88–99.
30. Hanson SJ, Halchenko YO (2008) Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Comput* 20:486–503.
31. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2430.
32. Sapountzis P, Schluppeck D, Bowtell R, Peirce JW (2010) A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. *Neuroimage* 49:1632–1640.
33. Rajimehr R, Young JC, Tootell RB (2009) An anterior temporal face patch in human cortex, predicted by macaque maps. *Proc Natl Acad Sci USA* 106:1995–2000.
34. Kriegeskorte N, Bandettini P (2007) Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage* 38:649–662.
35. Hadj-Bouziane F, Bell AH, Knusten TA, Ungerleider LG, Tootell RBH (2008) Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *Proc Natl Acad Sci USA* 105:5591–5596.
36. Leveroni CL, et al. (2000) Neural systems underlying the recognition of familiar and newly learned faces. *J Neurosci* 20:878–886.
37. Sugiura M, et al. (2001) Activation reduction in anterior temporal cortices during repeated recognition of faces of personal acquaintances. *Neuroimage* 13:877–890.
38. Damasio H, Tranel D, Grabowski T, Adolphs R, Damasio A (2004) Neural systems behind word and concept retrieval. *Cognition* 92:179–229.
39. Simmons WK, Reddish M, Bellgowan PS, Martin A (2010) The selectivity and functional connectivity of the anterior temporal lobes. *Cereb Cortex* 20:813–825.
40. Leopold DA, Bondar IV, Giese MA (2006) Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442:572–575.
41. Eifuku S, Nakata R, Sugimori M, Ono T, Tamura R (2010) Neural correlates of associative face memory in the anterior inferior temporal cortex of monkeys. *J Neurosci* 30:15085–15096.
42. Dehaene S, et al. (2010) How learning to read changes the cortical networks for vision and language. *Science* 330:1359–1364.
43. Hasson U, Levy I, Behrmann M, Hendler T, Malach R (2002) Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* 34:479–490.
44. Puce A, Allison T, Asgari M, Gore JC, McCarthy G (1996) Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *J Neurosci* 16:5205–5215.
45. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17:4302–4311.
46. Barton JJ, Press DZ, Keenan JP, O'Connor M (2002) Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology* 58:71–78.
47. Avidan G, Behrmann M (2009) Functional MRI reveals compromised neural integrity of the face processing network in congenital prosopagnosia. *Curr Biol* 19:1146–1150.
48. Thomas C, et al. (2009) Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. *Nat Neurosci* 12:29–31.

# Supporting Information

## Nestor et al. 10.1073/pnas.1102433108

### SI Text

**Stimulus Preparation.** *Faces.* Stimulus choice and construction were guided by two opposing goals. On the one hand, stimuli had to be as similar as possible with respect to a number of characteristics: high-level attributes (e.g., sex or age), low-level image descriptors (e.g., average luminance or contrast), and external feature properties (e.g., hair color or volume) to eliminate confounds with facial identity. On the other hand, individual faces needed to be as different from each other as possible to maximize the discriminability of the visually based activation patterns they elicit. To accommodate these different demands, we proceeded as follows.

First, we started with all front-view faces from the Face-Place 3.0 face database (www.face-place.org) and we narrowed down this dataset to young Caucasian adult male faces displaying a minimum of three basic emotional expressions (1) in addition to neutral expressions. This procedure ensures substantial within-identity image variability while preserving natural poses that are easy to interpret. In addition, we eliminated all faces that displayed facial hair, glasses, or other adornments, leaving us with a set of 128 faces (32 identities × 4 expressions).

Second, faces were normalized to the same size, subsampled to a lower resolution, and masked. More precisely, an oval mask was applied to all images to remove background and hair and also to reduce the dimensionality of the space (Fig. S6).

Third, we converted images to CIEL*a*b*, the color space that comes closest to that of human vision (2). Each image was normalized next with the same mean and contrast value separately for each of the three color channels: L* (corresponding to luminance), a* (corresponding to red:green), and b* (corresponding to yellow:blue).

Fourth, we computed pairwise similarity measures across all faces with a neutral expression. More specifically, we applied principal component analysis (PCA) to all faces and their mirror symmetric versions (3). We selected the projections on the first 40 principal components for each image and computed Mahalanobis distances between these lower-dimension patterns for each pair of neutral faces. A Mahalanobis metric was deployed given that it outperforms other types of metric with regard to both automatic face recognition (4) and modeling human face recognition (5). On the basis of these measurements, from all possible sets of four neutral faces we selected the set that minimized the average similarity score. We also ensured that each pair of faces within this set scored a similarity value below the average within the larger initial set.

Finally, we restored the original homogeneous background and applied the same hair feature to all four faces and their nonneutral versions (happy, sad, and disgusted). The resulting 16 images (Fig. 1) served as experimental stimuli for our individuation task.

A different set of faces was used for the functional localizers. ***Orthographic forms (OFs).*** Four five-letter pseudowords (Fig. S1) were presented in four different types of font (Arial Black, Comic Sans MS, Courier, and Lucida Handwriting). The pseudowords had the same syllable structure but were orthographically dissimilar in that they had no common letter in the same position. Moreover, they were composed of different sets of letters (with the exception of plang and greld that shared the letter "l" in different positions).

**Subjects.** Eight Caucasian young adults (five females, age range 18–22) from the Carnegie Mellon University community participated in the experiment. All subjects were right-handed and had normal or corrected-to-normal vision. None of the subjects had any history of neurological disorders. Two other subjects participated in the experiment; however, their data were excluded from analysis due to large head movements (more than a voxel) during at least one of three scanning sessions.

Informed consent was obtained from all subjects. The Institutional Review Board of Carnegie Mellon University approved all imaging and behavioral procedures.

**Behavioral Procedures.** Before scanning, subjects were presented with the 16-face stimuli described above and were trained to associate each facial identity with one of four buttons. None of the subjects were previously familiar with any of the faces presented nor were they given any biographical information with regard to them. Similarly, subjects were presented with the 16 OF stimuli and were trained to associate each individual OF with a button (face and OF responses were made using different hands randomly assigned to each category). Subjects practiced the task until accuracy reached ceiling (>98%). Training took place at least 1 d before each subject's first scanning session and was also briefly repeated before each scanning session.

During localizer scans, subjects performed a one-back task (same/different image). During the remaining functional scans, they performed the individuation task described above.

Stimuli were presented in the center of the screen against a black background and subtended a visual angle of 3.2° × 4.1°. Stimulus presentation and response recording relied on Matlab (Mathworks) and Psychtoolbox 3.0.8 (6, 7).

**Experimental Design.** Eight participants were each scanned for a total of 21 functional runs spread across three 1-h sessions. Of these, 17 runs used a slow event-related design whereas the rest used a block protocol suitable for functional localizers.

Localizer scans contained blocks of images grouped by category: faces, common objects, houses, words, and pseudofont strings. Each block consisted of back-to-back presentations of 15 stimuli for a total of 14 s (930 ms per stimulus). Stimulus blocks were separated by 10 s of fixation and were preceded by a 10-s fixation interval at the beginning of each run. No single stimulus was repeated within the course of a run. Each localizer scan contained 10 stimulus blocks, 2 for each stimulus category, and had a total duration of 250 s.

Runs with an individuation task used a slow event-related design with the following structure: a bright fixation cross was presented in the middle of the screen for 100 ms and then a stimulus appeared for 400 ms and was replaced by a lower-contrast fixation cross until the end of the event for 9.5 s. Each run contained a set of 32 such events following 10 s of fixation (for a total of 330 s). All face and OF stimuli described above were presented exactly once during each run. Stimuli were displayed in pseudorandom order to maximize uncertainty about stimulus identity (8) under the constraint that no more than three stimuli from the same category (face or OF) could be presented in a row.

Our decision to include OF stimuli along with faces was motivated by several different factors. First, inclusion of a different category was expected to reduce possible habituation/adaptation effects caused by prolonged exposure to the same small set of faces. Second, faces and OFs are perceptually highly dissimilar. Thus, although pattern discrimination for faces at the individual level is bound to be challenging for any type of method, discrimination of faces and OFs at the category level should be relatively easy and could serve as a robust benchmark for our classification method. Third and most important, the information

map obtained for face individuation could arguably be a generic individuation map, that is, not face specific but process specific. If so, we would expect other categories of objects with which we have extensive individuation experience, such as OFs, to produce similar information maps. Analysis of OF discriminability within the context of the same experiment provides us with a first test of this hypothesis. Finally, we opted for using pseudowords instead of actual words because they are unfamiliar (like faces) and minimize semantic processing while engaging similar mechanisms for OF processing (9, 10).

Functional scans were equally divided across three different sessions (seven scans per session) conducted on separate days. A structural scan was also performed at the beginning (or the end) of each session.

**Imaging Parameters.** Subjects were scanned in a Siemens Allegra 3T scanner with a single-channel head coil. Functional images were acquired with an echo-planar imaging sequence (TR 2 s, time to echo 31 ms, flip angle 79°, 2.5-mm isotropic voxels, field of view 240 × 240 $mm^2$, 27 oblique slices covering the ventral stream). An MP-RAGE sequence (1-$mm^3$ voxels; 192 slices of size 256 × 256 $mm^2$) was used for anatomical imaging.

**Preprocessing.** Functional scans were slice scan time corrected, motion corrected, coregistered to the same anatomical image, and normalized to percentage of signal change using AFNI (11). Functional localizer data were smoothed with a Gaussian kernel of 7.5 mm FWHM. No spatial smoothing was performed on the rest of the data to allow multivariate analysis to exploit high-frequency information (12).

**Standard Univariate Analysis.** After completion of preprocessing steps we discarded the first 5 vol of each run to allow the hemodynamics to achieve a steady state and to minimize transient effects of magnetic saturation. Next, we fitted each type of block with a boxcar predictor and convolved it with a gamma hemodynamic response function (13). A general linear model (14) was applied to estimate the coefficient of each predictor independently for each voxel. Statistical maps were computed by $t$ tests of pairwise comparisons between different block types. Face-selective areas were detected using a face–object contrast. Correction for multiple comparisons was implemented by controlling the false discovery rate under the assumption of positive/no correlation (15).

**Spatiotemporal Information-Based Brain Mapping.** A manually drawn cortical mask was constructed for each subject's brain. Fig. S2*A* shows the corresponding group mask. Searchlight analysis was carried out by walking a sphere voxel-by-voxel across the entire volume of the mask, extracting the spatial–temporal patterns recorded at each location, and testing them for the presence of relevant information via multivariate analysis. More specifically, a sphere with a 5-voxel radius was centered on each voxel within the cortical mask and intersected with the mask to restrict analysis to cortical voxels. Activation values across this restricted set of voxels at three different time points (4, 6, and 8 s after stimulus onset) were extracted for each stimulus presentation and concatenated into a single pattern.

Our choice of a 5-voxel spatial radius was based on pilot data not included in the current analysis. In addition, to test the sensitivity of our results as a function of this parameter, we conducted identical analyses for searchlight radii of 4 and 6 voxels. We note that increasing the size of the searchlight may both benefit and hurt the mapping results and their interpretation. A larger searchlight augments the amount of potentially useful information but also increases the dimensionality of the patterns leading to more overfitting. Also, the larger the searchlight is, the less local the mapping results will be: Highly local information will be exploited by all searchlight masks that contain it over a larger area, thus leading to a more diffuse map—see Fig. S3 for an example. Our choice represents a compromise between searching for local information and exploiting a sufficient amount of spatial information.

The temporal size of the window was selected to capture the peak of the hemodynamic response function (HRF) (16). We note that a full-blown version of a spatial–temporal searchlight would have to walk a window in both space and time. Whereas this approach may provide a more detailed assessment of the temporal–spatial profile of information maps, such analysis comes at significant additional computational cost. As an alternative to this approach, we restrict our analysis to spatial mapping and keep the position of our temporal window fixed.

Next, to boost the signal-to-noise ratio (SNR) of our patterns, we averaged stimulus-specific patterns by stimulus identity. Thus, all patterns elicited during a functional run by images of the same individual, irrespective of the expression displayed, were combined into a single one. This procedure produced 17 different patterns, 1 per run, for each of four different facial identities. A similar procedure was used for OF stimuli.

To measure identity discriminability, we applied multiclass SVM classification using a one-against-one approach to speed up computations (17)—that is, each facial identity is compared with every other one at a time. Our particular choice of classifier is linear SVM with a trainable $c$ term because it appears to perform better or, at least, equivalently to other classifiers tested on neuroimaging data (18, 19). Leave-one-run-out cross-validation was carried out for each pair of facial identities. At the same time, nested cross-validation within each training set was conducted to optimize the $c$ term (allowed to range between $2^{-4}$ and $2^{10}$) and minimize overfitting. Discriminability was next encoded using the sensitivity measure $d'$ (20). Voxelwise averaging of these estimates across each of the six pairs of identities compared produced subject-specific information maps.

Because it appears that multivariate analysis is able exploit high-frequency spatial information (12, 21), we attempted to minimize the amount of distortion of the functional data and preserve this information. Thus, multivariate analysis was carried out on unsmoothed data in each subject's native space. However, for the purpose of group analysis all information maps were brought into Talairach space. Group information maps were obtained by averaging across subjects and statistical effects were computed using a one-sample $t$ test against chance ($d' = 0$). Finally, multiple-comparison correction was implemented using FDR.

Whereas the analysis described above was designed to take advantage of information distributed across patterns of activation, it is possible that patterns per se contribute little, if anything, to the effects detected. In other words, it is possible that multivariate effects present in the data can be accounted for by univariate effects. To test this hypothesis, we carried out an analysis following the same procedure with the sole difference that patterns are averaged into single values previous to classification. This simplification renders the analysis comparable to a univariate $t$ test.

Finally, we conducted a similar set of multivariate analyses to examine expression discrimination and OF individuation as well as category-level classification (faces versus OFs). More precisely, we computed discrimination performance among (*i*) four different expressions across changes in facial identity, (*ii*) four different OF identities across changes in font type, and (*iii*) two categories across variations in both identity and category-specific changes. In all other respects, the computation of the respective information maps follows the procedure described above.

With respect to category-level discrimination, we note two factors that need to be taken into account. First, the categories being discriminated, faces and OFs, are very dissimilar, both perceptually and conceptually. Second, the SL size was larger than that typically used, e.g., a 2-voxel radius (22), and represented a compromise between maintaining a local encoding con-

straint and maximizing the amount of spatial information as discussed above. Thus, it is possible that category information is somewhat more focal than we ended up finding (Fig. S2). Nevertheless, category differentiation was sufficiently dispersed to produce a rather diffuse information-based map.

Analyses were carried out in Matlab using the SVMLIB 2.88 library for pattern classification (23).

**Note on the Use of Spatiotemporal Information in Pattern Classification.** The use of spatiotemporal information for multivariate analyses (24) presents us with an interesting opportunity. The temporal properties of the BOLD signal (e.g., time to peak or time to rise) provide a rich source of information regarding the neural dynamics (25). However, both their interpretation in relationship with the actual neural dynamics and their estimation can be problematic (26)—although not more so than that of the ubiquitously used signal amplitude. Multivariate spatiotemporal analysis (24, 27) allows us to bypass the latter problem in that no estimation of temporal properties (or amplitudes for that matter) is required. Rather, the use of such information is implicit and, thus, eliminates the issue of model (mis)specification (26).

Furthermore, if selecting a single time point for the analysis, it is unclear which one encodes the most diagnostic spatial information. The HRF peak may lead, in certain cases, to the best decoding accuracy (28). However, the shape of the HRF, including the timing of the peak, can vary significantly among cortical areas and across subjects (29). Fortunately, the use of multiple time points allows for the possibility that diagnostic spatial information, whether corresponding to the response peak or not, can be present at different times in different areas or across different subjects. Our analysis above uses a combination of these two approaches by exploiting multiple time points while restricting their number to those likely to capture the HRF peak.

Overall, the advantages above make spatiotemporal analysis very appealing as long as the increase in pattern dimensionality introduced by this approach can be handled adequately (e.g., by the use of classifiers that scale well with dimensionality).

**RFE Analysis.** RFE serves three different but related goals: dimensionality reduction within the original feature space, optimization of classification performance, and feature ranking (30). The analysis proceeds as follows: (*i*) we train a linear SVM classifier on a given feature set, (*ii*) we compute a ranking criterion for all features, (*iii*) we eliminate the feature with the smallest rank, and (*iv*) we repeat until no features are left in the set.

One of the simplest feature ranking criteria for linear SVM, and the one we follow here, is based on maximizing the separating margin width of the classifier (30). More specifically, the algorithm eliminates within each iteration the feature with the smallest $c_i = w_i^2$, where $w_i$ is the weight corresponding to feature $i$. This procedure has the effect of maintaining the largest possible margin width $W = \|w\|$ at each iteration step. The number of iteration steps corresponds, in this version of the algorithm, to the total number of features in the initial set. Whereas batch elimination provides an easy alternative to speeding up computations, it may lead to suboptimal estimates of performance and also compromise feature ranking. For this reason, we favored single-feature elimination in our analysis.

RFE analysis has been successfully used in the past to reduce the dimensionality of fMRI data (31) and to map voxel diagnosticity in category-level discrimination (32). Here, we use it both to map feature (voxel $X$ time point) diagnosticity and to improve on the classification models derived for individuation.

The analysis was separately applied to each pair of facial identities. Diagnosticity rankings as well as performance estimates were computed by averaging across all six different pairs.

**Contribution of Low-Level Image (Dis)Similarity to Individuation Performance.** Low-level similarity was computed between any two images of different facial identities. More precisely, we applied an $L_2$ (Euclidian) metric (4) to estimate low-level image dissimilarity. To ensure the robustness of the results, the metric was applied in three different ways corresponding to different ways of extracting information: (*i*) to entire images using only the luminance channel, (*ii*) to cropped images using only luminance, and (*iii*) to cropped images using all color channels (Fig. S6).

The manipulations above are motivated, first, by the privileged role of internal features relative to external ones in face perception (33) and, second, by the contribution of color to face processing (34, 35). Thus, to deal with the first issue we used cropping to eliminate external features (e.g., hair and face outline) and to retain internal ones (e.g., eyes and mouth). To deal with the second, we combined similarity measures computed independently for each color channel. However, whereas color is known to be involved both in low-level (36) and in high-level face processing (35), the relative contribution of different color channels is still unclear (37). For this reason, all channels were given equal weight in computing the estimates corresponding to case *iii* above. Specifically, channel-specific estimates were $z$-scored across image pairs and then averaged to produce single values for each pair. Finally, the values obtained for all 16 image pairs corresponding to two different identities were averaged to produce a single score (Table S1).

The three types of measurement are in overall agreement with each other: identities 1 and 2 along with 2 and 4 are relatively similar to each other whereas 3 and 4 are the most dissimilar (identity numbers refer to the columns in Fig. 1).

Next, dissimilarity estimates were correlated with individuation performance across identity pairs separately for each ROI, experimental subject, and type of measurement. The resulting correlation coefficients were converted into normally distributed variables using Fisher's $z$ transform, allowing us to conduct parametric tests on the results:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Finally, average subject scores were compared against chance via one-group $t$ statistics (Fig. S4).

**Pairwise ROI Analysis.** ROI-based classification patterns provide only a coarse and summary measure of the relevant information present in the ROIs, namely in the activation patterns they host. However, they can be useful in that they offer an estimate of common biases in misclassification.

To compare classification patterns produced by different ROIs we used both partial correlation and conditional mutual information while controlling for the pattern of correct (true) labels.

Correlation coefficients were converted to $z$ scores using Fisher's $z$ transform.

Conditional mutual information (38) was computed as follows:

$$I(C_1, C_2|T) = \sum_{\substack{C_1,C_2\in\{0,1\}\\ T\in\{0,1\}}} p(C_1, C_2, T)\log\left(\frac{p(C_1, C_2|T)}{p(C_1|T)p(C_2|T)}\right).$$

Here $C_1$ and $C_2$ are binary variables encoding the classification labels for two different regions and $T$ is a binary variable encoding the true labels.

The two measures were separately computed for each pair of ROIs and averaged across face pairs and subjects.

1. Eckman EP (1972) *Nebraska Symposium on Motivation*, ed Cole CJ (Univ of Nebraska Press, Lincoln, NE), pp 207–283.
2. Brainard DH (2003) *The Science of Color*, ed Shevell SK (Optical Society of America, Washington, DC), pp 191–216.
3. Turk M, Pentland A (1991) Eigenfaces for recognition. *J Cogn Neurosci* 3:71–86.
4. Moon H, Phillips PJ (2001) Computational and performance aspects of PCA-based face-recognition algorithms. *Perception* 30:303–321.
5. Burton AM, Miller P, Bruce V, Hancock PJ, Henderson Z (2001) Human and automatic face recognition: A comparison across image formats. *Vision Res* 41:3185–3195.
6. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10:437–442.
7. Brainard DH (1997) The psychophysics toolbox. *Spat Vis* 10:433–436.
8. Wager TD, Nichols TE (2003) Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *Neuroimage* 18:293–309.
9. Vigneau M, Jobard G, Mazoyer B, Tzourio-Mazoyer N (2005) Word and non-word reading: What role for the Visual Word Form Area? *Neuroimage* 27:694–705.
10. Polk TA, Farah MJ (2002) Functional MRI evidence for an abstract, not perceptual, word-form area. *J Exp Psychol Gen* 131:65–72.
11. Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29:162–173.
12. Swisher JD, et al. (2010) Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. *J Neurosci* 30:325–330.
13. Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* 16:4207–4221.
14. Friston KJ, et al. (1995) Statistical parametric maps in functional imaging: A general linear approach. *Hum Brain Mapp* 2:189–210.
15. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
16. Friston KJ, Jezzard P, Turner R (1994) Analysis of functional MRI time-series. *Hum Brain Mapp* 1:153–171.
17. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw* 13:415–425.
18. Pereira F, Botvinick M (2011) Information mapping with pattern classifiers: A comparative study. *Neuroimage* 56:476–496.
19. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* 53: 103–118.
20. Salton G, McGill M (1983) *Introduction to Modern Information Retrieval* (McGraw-Hill, New York).
21. Kriegeskorte N, Cusack R, Bandettini P (2010) How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? *Neuroimage* 49:1965–1976.
22. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
23. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2:27.
24. Mourão-Miranda J, Friston KJ, Brammer M (2007) Dynamic discrimination analysis: A spatial-temporal SVM. *Neuroimage* 36:88–99.
25. Formisano E, Goebel R (2003) Tracking cognitive processes with functional MRI mental chronometry. *Curr Opin Neurobiol* 13:174–181.
26. Lindquist MA, Meng Loh J, Atlas LY, Wager TD (2009) Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *Neuroimage* 45(1, Suppl): S187–S198.
27. Mitchell TM, et al. (2004) Learning to decode cognitive states from brain images. *Mach Learn* 57:145–175.
28. Johnson JD, McDuff SG, Rugg MD, Norman KA (2009) Recollection, familiarity, and cortical reinstatement: A multivoxel pattern analysis. *Neuron* 63:697–708.
29. Handwerker DA, Ollinger JM, D'Esposito M (2004) Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* 21:1639–1651.
30. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46:389–422.
31. De Martino F, et al. (2008) Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43:44–58.
32. Hanson SJ, Halchenko YO (2008) Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Comput* 20:486–503.
33. Andrews TJ, Davies-Thompson J, Kingstone A, Young AW (2010) Internal and external features of the face are represented holistically in face-selective regions of visual cortex. *J Neurosci* 30:3544–3552.
34. Bindemann M, Burton AM (2009) The role of color in human face detection. *Cogn Sci* 33:1144–1156.
35. Nestor A, Tarr MJ (2008) Gender recognition of human faces using color. *Psychol Sci* 19:1242–1246.
36. Yip AW, Sinha P (2002) Contribution of color to face recognition. *Perception* 31: 995–1003.
37. Nestor A, Tarr MJ (2008) The segmental structure of faces and its use in gender recognition. *J Vis* 8:7–.1–12.
38. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).



**Fig. S1.** Experimental orthographic form (OF) stimuli (four pseudowords × four types of font). All stimuli were five-letter pronounceable nonwords with the same syllabic structure but different orthographic properties (they contain different letters in a given position).

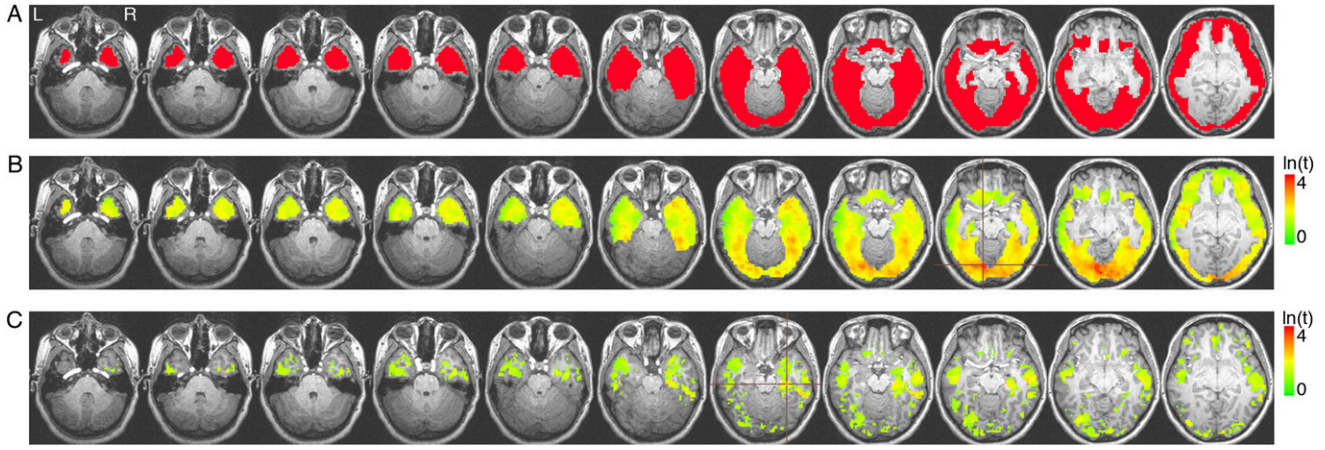


**Fig. S2.** (*A*) Group cortical mask and group information-based maps of category-level discrimination (faces vs. OFs; $q < 0.05$) derived through (*B*) multivariate searchlight and (*C*) its univariate analog. Effect size is scaled logarithmically. Crosshairs mark the discrimination peaks in each map (Talairach coordinates −11, −76, −11 and 31, −24, −16 for *B* and *C*, respectively). The differences between the two types of map indicate that univariate analysis underestimates the amount and expanse of category information in ventral cortex compared with its multivariate counterpart.

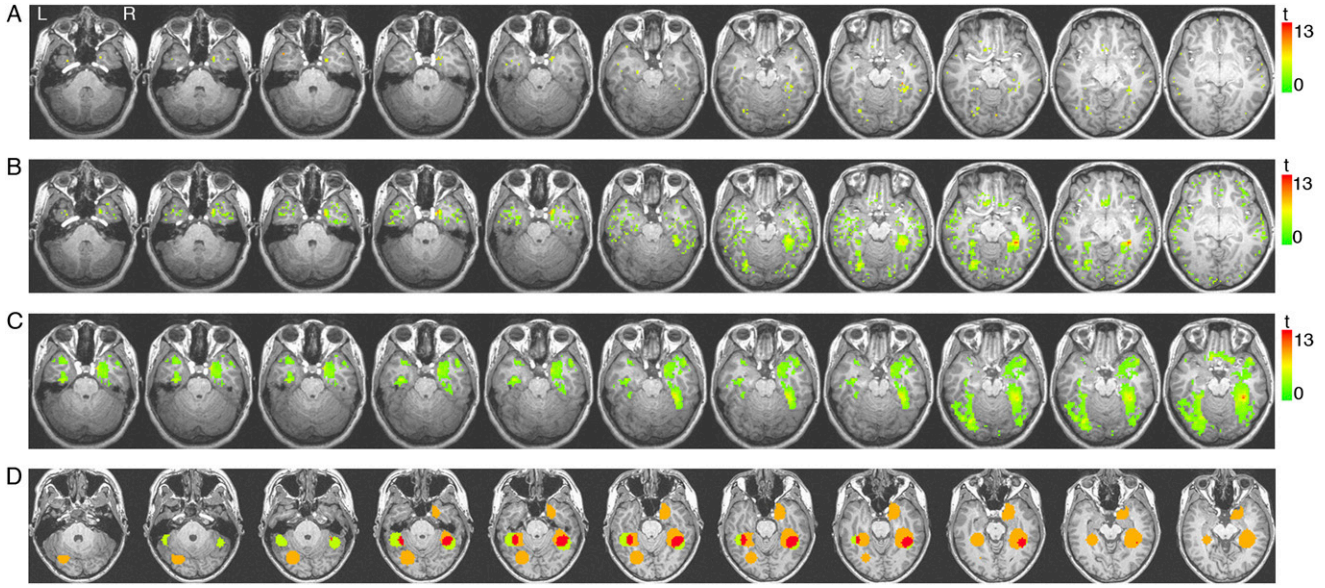


**Fig. S3.** (*A–C*) Group information-based maps of face individuation obtained with SL masks of different spatial radii (4, 5, and 6 voxels). Maps were thresholded at a liberal level ($q < 0.10$) for better comparison with each other. Larger masks yielded a larger number of significant voxels as expected on the basis of the loss of spatial specificity. Despite such variation, the three maps agree on the presence of individuation effects in the bilateral FG and the right aMTG. (*D*) Examples of SL ROIs (orange), bilateral FFA (green), and their overlap (red) within a single subject.

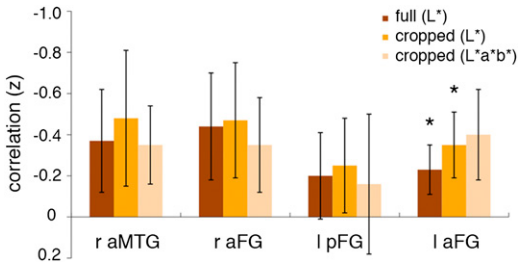


**Fig. S4.** Correlation of discrimination performance with low-level image distances across different SL ROIs (Fisher's *z* scores). Low-level distances were computed in three different ways: across the entire image using only luminance (full L*), across internal features using only luminance (cropped L*), and across internal features using all color channels (cropped L*a*b*). The laFG produced correlations of discrimination performance and image distance for luminance-based measurements ($P < 0.05$ uncorrected for multiple comparisons)—the effect was less reliable across all color channels ($P < 0.06$ uncorrected).

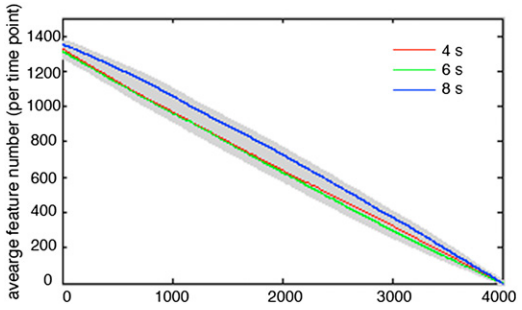


**Fig. S5.** The time course of feature elimination by time point for 4,000 features—top 1,000 features for each of four ROIs. Features tend to be eliminated at similar rates across different time points although we note a small advantage for features extracted at 8 s poststimulus onset (shaded areas correspond to ±1 SE across subjects).

**Fig. S6.** Examples of image cropping used for stimulus selection and for measurements of low-level image similarity. Images were cropped to show only internal features. Face images courtesy of the Face-Place Face Database Project (http://www.face-place.org/) Copyright 2008, Michael J. Tarr. Funding provided by NSF Award 0339122.

**Table S1. Estimates of low-level image distances across facial identity pairs (*z* scores)**

| Identity pair | Full image (L*) | Cropped image (L*) | Cropped image (L*a*b*) |
|---|---|---|---|
| 1–2 | −0.31 | −0.54 | −1.07 |
| 1–3 | −0.19 | 0.22 | 0.31 |
| 1–4 | 0.58 | 0.08 | 0.04 |
| 2–3 | −0.15 | 0.13 | −0.08 |
| 2–4 | −0.91 | −0.91 | −0.38 |
| 3–4 | 0.98 | 1.02 | 1.19 |

OPEN ACCESS Freely available online

PLoS one

# Task-Specific Codes for Face Recognition: How they Shape the Neural Representation of Features for Detection and Individuation

**Adrian Nestor**[9], **Jean M. Vettel**[9], **Michael J. Tarr**[9]*

Department of Cognitive and Linguistic Sciences, Brown University, Providence, Rhode Island, United States of America

## Abstract

***Background:*** The variety of ways in which faces are categorized makes face recognition challenging for both synthetic and biological vision systems. Here we focus on two face processing tasks, detection and individuation, and explore whether differences in task demands lead to differences both in the features most effective for automatic recognition and in the featural codes recruited by neural processing.

***Methodology/Principal Findings:*** Our study appeals to a computational framework characterizing the features representing object categories as sets of overlapping image fragments. Within this framework, we assess the extent to which task-relevant information differs across image fragments. Based on objective differences we find among task-specific representations, we test the sensitivity of the human visual system to these different face descriptions independently of one another. Both behavior and functional magnetic resonance imaging reveal effects elicited by objective task-specific levels of information. Behaviorally, recognition performance with image fragments improves with increasing task-specific information carried by different face fragments. Neurally, this sensitivity to the two tasks manifests as differential localization of neural responses across the ventral visual pathway. Fragments diagnostic for detection evoke larger neural responses than non-diagnostic ones in the right posterior fusiform gyrus and bilaterally in the inferior occipital gyrus. In contrast, fragments diagnostic for individuation evoke larger responses than non-diagnostic ones in the anterior inferior temporal gyrus. Finally, for individuation only, pattern analysis reveals sensitivity to task-specific information within the right ''fusiform face area''.

***Conclusions/Significance:*** Our results demonstrate: 1) information diagnostic for face detection and individuation is roughly separable; 2) the human visual system is independently sensitive to both types of information; 3) neural responses differ according to the type of task-relevant information considered. More generally, these findings provide evidence for the computational utility and the neural validity of fragment-based visual representation and recognition.

**Citation:** Nestor A, Vettel JM, Tarr MJ (2008) Task-Specific Codes for Face Recognition: How they Shape the Neural Representation of Features for Detection and Individuation. PLoS ONE 3(12): e3978. doi:10.1371/journal.pone.0003978

**Editor:** Ernest Greene, University of Southern California, United States of America

**Received** September 12, 2008; **Accepted** November 18, 2008; **Published** December 29, 2008

**Copyright:** © 2008 Nestor et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This research was supported by NSF Award #0339122, by the Temporal Dynamics of Learning Center (NSF Science of Learning Center SBE-0542013), and through the generosity of The Ittleson Foundation. The MRI system used in the study was purchased in part by an MRI grant from the NSF. JMV was supported by a DOD SMART Graduate Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: michael_tarr@brown.edu

[9] These authors contributed equally to this work.

## Introduction

One of the hallmarks of human face processing is our ability to recognize faces across a multitude of levels. At the most general level, we are able to locate and distinguish faces from non-faces in a visual scene. We can also categorize faces by any number of traits including gender, ethnicity, expression, and age. Finally, we routinely discriminate faces from one another at the individual level (which we refer to as ''individuation''). This range of abilities leads us to ask whether the visual system carries out such categorization tasks using a single general type of category representation or, alternatively, translates task-specific constraints into different representations of the overall category. Our study focuses on the two ends of this categorization spectrum by comparing the sensitivity of the human visual system to face detection and face individuation.

Artificial systems for automatic face recognition typically treat detection and individuation as two separate problems with two different goals [1]. However it is less clear to what extent biological systems such as the human visual system adopt a similar dual approach. Models of human face processing acknowledge the difference between multiple face recognition tasks. For instance, two classical models of face processing [2] and its neural basis [3] are centered around task differences. However, the main dichotomy these models emphasize is that between expression recognition and individuation. While an early stage of facial feature processing is separated from these tasks, the locus of face detection as well as its relationship with this early stage and the

other tasks are less spelled out. Neuroimaging results also provide mixed evidence for the neural separability of detection and individuation. Some studies found that a set of common areas in the fusiform gyrus [4–7] and the inferior occipital gyrus [5,7] are sensitive to both face detection and individuation. However, a recent study [8] uncovered an area sensitive to face individuation in the right anterior inferior temporal cortex, outside the typical face-selective regions recruited by detection. Consistent with this result, studies of white matter connectivity linked behavioral individuation performance with the structural integrity of fibers connecting the fusiform gyrus with more anterior areas in the right hemisphere, including the inferior temporal cortex [9,10]. Finally, the neural markers of the time course for these two processes also seem to be different: detection and individuation were associated with separate M100 and M170 components in a magnetoencelography study [11].

If face detection and individuation do recruit different brain areas and exhibit different time courses, this may point to processing and representational differences that characterize and motivate their separation. One possibility is that the two tasks we consider here pose objectively different constraints on the featural codes underlying these two types of face recognition. The present study investigates the neural separability of detection and individuation precisely by exploring this issue. We hypothesize that detection and individuation require separate sets of facial features to optimally achieve their goals and that the visual system adopts this separation to perform different aspects of face recognition more effectively. Moreover, we surmise this difference in the representational bases of the two tasks leads in turn to differences in neural processing sufficiently robust to be tested by functional magnetic resonance imaging (fMRI).

The present study investigates this hypothesis by appealing to a framework for synthetic vision initially developed in the context of automatic object detection [12]. More recently this framework has been extended to a model of human vision and has been evaluated with respect to its original use as a method for category detection [13,14] and category learning [15]. Within this framework, categories of objects are represented as sets of overlapping rectangular image fragments of different aspect ratios, sizes and resolutions. Many other candidates for the role of object and facial features have been proposed in the literature: edge structures [16,17], principal components of images [18,19] or image segments [20,21], to name just a few. However, for the goals of our present study, we find fragment features appealing for a number of reasons. First, they are cue-agnostic in that they do not commit themselves from the start to a single type of cue, for example, edges. Second, they naturally account for configural information, an important aspect of face processing [22], in that allowing features to overlap constrains the spatial relationship of otherwise disjoint image fragments. Finally, and most relevant for the objectives of our present study, Ullman et al's framework provides a principled means for establishing optimal *task-specific* sets of fragment features for a given category. In the case of faces, it has been shown, for instance, that such features lend themselves naturally to deal not only with detection [12] but also with other categorization tasks, for example, individuation or expression recognition [23].

We note the problem of mapping diagnostic areas for specific object categories and particular tasks has been investigated in human observers using other approaches. For instance, reverse correlation methods [24–26] and 'bubbles' [27,28] in particular, can effectively produce maps of task-diagnostic regions of images with respect to a visually-homogeneous category such as faces. Critically, the task-diagnostic maps produced by these methods are compatible with different models of how one might divide an object into features. The fragment-based approach we adopt here produces concrete ways to decompose a stimulus category into feature components. We take advantage of this property to gain a finer-grained view of the representational codes used in task-specific face processing.

Our investigation proceeds in three stages. First, we evaluate and compare systematically the task-specific information of face fragments for detection and individuation. This evaluation enables the selection of sets of fragments whose task-specific information varies independently in the two tasks. Second, using these fragments, we assess and confirm the sensitivity of human subjects to task-diagnostic fragment features. Third and finally, an fMRI study tests and reveals that different cortical areas exhibit different patterns of sensitivity to task-specific information with respect to our two tasks. These results jointly confirm the separability of detection and individuation in the human visual system and provide evidence for different representational codes underlying the two tasks and driving the noted separation.

## Materials and Methods

### Evaluation of Task-Specific Information

**Stimuli.** Image face fragments were extracted from a set of 60 face images (12 individuals×5 expressions)–Figure S1–selected from the Tarrlab face database (available online at www.face-place.org). This set contains near-front-view grayscale Caucasian faces, half of which were male and half female, wearing no glasses or other facial accessories and displaying variable affective expressions. The faces were cropped, down-sampled (60×40 pixels) and normalized with the position of the main features, the eyes and the nose, to permit the mapping of corresponding face fragments across faces. A similar set containing 12 different individuals was used for cross-validation of the computational results. In addition, two sets of 605 natural scene images were randomly selected from the McGill Calibrated Color Image Database (tabby.vision.mcgill.ca). These images, mapped to grayscale, were used in the computation of the amount of detection-specific information and also provided non-face stimuli for behavioral testing.

The image fragments were rectangular image patches of different sizes and aspect ratios [12]. More precisely, we systematically varied the size, aspect ratio and position of a rectangular window across each face in steps of 4 pixels. Thus, the smallest fragments were 4×4 pixel image patches while the largest ones contained an entire face. For each face this procedure yielded 6600 image fragments. Consequently, application of the same procedure to all face images yielded 6600 *fragment types*, where by fragment type we mean a class of fragments corresponding to the same area of the face across different images of the same or different individuals. Examples of face fragments extracted from the same image are shown in Figure 1.

**Methods.** The amount of face-detection information was computed for each of 396000 fragments generated by extracting 6600 rectangular fragment types from 60 face images. We refer to these fragments as well as the face images from which they were extracted as the training set. A separate test set was composed of an equal number of fragments extracted from a different set of face images.

To estimate the task-specific information of a fragment of a given type $k$, we cross-correlated the given fragment with all face and non-face images. If the maximum correlation value surpassed a certain threshold $\theta_k$, the fragment was considered present in the image. The threshold $\theta_k$ was computed for each fragment type $k$ so

**Figure 1. Example of face fragments extracted from the same face (displayed in reduced contrast).**
doi:10.1371/journal.pone.0003978.g001

as to maximize the average task-specific information of the fragments of this type for face detection. Computation of the amount of task-specific information carried by each fragment was implemented following the original description of the method [12]. Briefly, for each face fragment we computed these values using the mutual information between fragment presence and image category, that is, face or non-face. The mutual information was computed as [29]:

$$I(F,C)=\sum_{\substack{F=\{0,1\}\\C=\{0,1\}}} p(F,C)\log\left(\frac{p(F,C)}{p(F)p(C)}\right)$$

Here F is a binary variable indicating whether a given fragment was present in an image or not and C is a binary variable indicating whether the image contains a face or not. The threshold $\theta_k$ was estimated by maximizing the mutual information of a fragment over the training set (the best threshold was found by brute-force search from $-0.99$ to 0.99 in steps of 0.01). In a departure from the original method, a common threshold was estimated for each fragment type rather than for each fragment separately. The overall task-specific information for a fragment type was computed as the average mutual information of all fragments of that type in the training set.

The method described above was next extended to individuation. For this task the training set was limited to faces and C denotes in this case same/different individuals instead of face/non-face information. More precisely, C encodes whether an image contains the face of the same individual from which the fragment tested was extracted or not–for any particular fragment there are 4 different images of the same individual in addition to the one from which the fragment was initially extracted and another 55 faces of 11 different individuals. A new set of task-specific thresholds for fragment presence was estimated again for each fragment type.

We note that a given face fragment can fail to be informative for detection because it is not similar enough to other fragments of the same type, because it is highly similar to recurrent image structures found in non-face images or because of both. (Figure S2 shows natural image fragments erroneously labeled as face fragments by the method due to their similarity to actual face fragments.) It is possible a fragment is highly diagnostic of a particular individual but, due to the high variability of the type to which it belongs across different individuals, it is less useful for detection. Thus, in order to be diagnostic for the two tasks, fragment types need to satisfy two different criteria: small *extrapersonal* (between-individual) variability for detection versus large *extrapersonal* variability relative to *intrapersonal* (within-individual) variability for individuation [30]. If the two criteria conflict for most fragments, we would expect a relatively low correlation between their task-specific information values for the two tasks. To verify this hypothesis, we computed the Pearson correlation between the mutual information for detection and individuation across all fragment types.

For cross-validation purposes, task-specific information for each fragment type as well as the correlation between values for detection-specific and individuation-specific information were computed again within the test set. Cross-correlation thresholds in this case were kept fixed at the values that maximized task-specific information within the training set.

## Behavioral Experiment 1–Face Detection

**Participants.** Sixteen adults from the Brown University community volunteered to participate in the experiment in exchange for pay. All participants had normal or corrected-to-normal vision. All participants provided written consents and procedures were approved by the Institutional Review Board of Brown University.

**Stimuli.** From our 6600 face fragment types, we preselected a subset adequate for the testing of human participants. For every fragment type we verified whether it contained any subfragments with the value of detection-specific information higher or equal to that of its own. If this was not the case, the fragment type in question was included in the mentioned subset. The fact that the overall task-specific information for a fragment can be lower than that of a subfragment it contains owes to the fact that the evaluation of its task-specific information weighs in equally all pixels. The selection criterion imposed above ensures the stimuli used are categorized correctly with respect to the amount of information they provide to our observers. In its absence, overall task-specific information may be misleading in studying human vision in that participants can zero in on the most diagnostic subfragment(s) of a given fragment and disregard the rest.

Next, forty fragment types were selected for each of three levels of detection-specific information (high, middle and low) from our preselected set of candidates. Fragment types were selected so as to homogenize the entire set of stimuli with respect to potential confounds. Task-specific information for the irrelevant task, that is, individuation, as well as geometric properties, size (in number of pixels) and aspect ratio, were all considered in selecting the final set of stimuli ($p>0.1$ for all pairwise comparisons between the different levels of task-specific information across irrelevant dimensions). Finally, for each of the forty fragment types, the actual stimuli were selected by picking two fragments of that type from two randomly selected faces of two different individuals. In addition, 240 natural image fragments were extracted to match the face fragments with respect to their geometric properties. Contrast and mean luminance was equalized across all face and natural image fragments.

We note that the qualitative labels applied to the three levels of task-specific information are meaningful relative to each other rather than by absolute ranking with respect to ideally diagnostic fragments–see Table 1. This is not a reason of concern for

recognition performance in that recognition, in this case detection, does not rely on single ideal features but on multiple features that jointly contribute to the process possibly based on their independent amounts of information they provide [12].

**Task.** Participants were asked to perform a face detection task by pressing one of two buttons on a buttonbox. More precisely, participants were asked to judge whether the single image fragment displayed at a time was a face fragment or not. The response was made by pushing one of two buttons with the index fingers of the left and right hands randomly assigned to signal a face/non-face response across participants.

On each trial, a cross was presented in the center of the screen for 400 ms followed by an image fragment for 250 ms. A black screen replaced the stimulus until the participant made a response signaling the end of a trial. A stimulus subtended on the average a visual angle of $2.1\times2.3$ from a distance of 70 cm after doubling the size of the image by pixel replication. Each participant completed 480 trials over the course of two blocks in a 30-minute session. Each stimulus was shown only once in the entire experimental session. Trial order was randomized for each participant.

Experimental trials were preceded by a short practice session allowing the participants to familiarize with the task and the stimuli. At the end of the experiment participants were asked to report whether they were familiar with and able to recognize any of the individuals whose faces were presented in the experiment.

Stimulus design and presentation relied on Matlab 7.5 (Mathworks, Natick, MA, USA) and the Psychophysics Toolbox 3 [31,32] running on an Apple Macintosh using OS X.

## Behavioral Experiment 2–Face Individuation

**Participants.** Another sixteen adults from the Brown University community with normal or corrected-to-normal vision participated in the experiment. All participants provided written consents and procedures were approved by the Institutional Review Board of Brown University.

**Stimuli.** From all our face fragment types, we preselected a subset in a manner analogous to the one described for the first experiment. However, this time the relevant task-specific information was computed for individuation instead of detection.

Next, the procedure described above was followed to select 40 fragment types for each of three levels of individuation-specific information. In addition to controlling for low-level properties of the stimuli we also attempted to homogenize the overall set with respect to detection-specific information–see Table 1. Finally, for each of the forty fragment types, the actual stimuli were selected by picking four fragments of that type from two individuals where each of these individuals supplied two distinct fragments of that type showing two different expressions.

Interestingly, we note that, as a result of our selection procedure, the average size of a fragment in this experiment was significantly larger than the one in the first experiment (two-tailed $t$-test $p<0.01$). This difference is mainly due to the fact that it is more difficult to find intermediate-sized fragments with small detection-specific information than it is to find small fragments with high task-specific information. Conversely, it is more difficult to find small fragments with high individuation-specific information than to find intermediate-sized fragments with low amounts of relevant information.

**Task.** Participants were asked to judge whether two fragments *of the same type* shown in succession belonged to the same individual or not. Each trial had the following structure: a cross appeared for 400 ms in the center of the screen followed by the first image fragment for 250 ms, a white noise mask with the same size as the previously presented image for 200 ms, the second face fragment for another 250 ms and a black screen until the subject made a button press. Trials were equally divided between same-individual versus different–individual trials for each condition. A stimulus subtended on the average a visual angle of $3.4\times2.3$ from a distance of 70 cm after doubling the size of the image. Each participant completed 240 trials over the course of two blocks in a single a 45-minute session. Each stimulus was shown only once in the entire experimental session and trial order was randomized for each participant. In all other respects we followed the procedure described for Experiment 1.

## Functional Magnetic Resonance Imaging (fMRI) Experiment

**Participants.** Eleven healthy adult members (7 female, age range: 18–30) of the Brown University community, including one of the authors JMV, volunteered to participate in the experiment for pay. None of them took part in the behavioral experiments described above. Participants were right-handed, with normal or corrected-to-normal vision and reported no contraindications for MRI scanning. All participants provided written consents and procedures were approved by the Institutional Review Board of Brown University.

**Stimuli and behavioral task.** Participants were presented with the same face fragment stimuli as the ones from the behavioral experiments but using only two levels of task-specific information: high and low. Participants lay supine and viewed the rear-projection display through an angled mirror in the bore of the magnet. Stimuli were presented in 30-second blocks of face fragments with high/low levels of information for detection or individuation. The order of the blocks was counterbalanced across participants. Half of the stimuli in each condition were presented twice during the experiment but at most once within a block. Stimulus duration was 800 ms with 700 ms inter-stimulus interval (20 stimuli per block). The stimuli in the detection and individuation conditions subtended different viewing angles similar to those in the behavioral experiment. In each time-

**Table 1.** Task-specific information* carried by fragments used in the two behavioral tasks.

| Level of Information | Detection task | | Individuation task | |
|---|---|---|---|---|
| | detection MI | individuation MI | detection MI | individuation MI |
| high | 0.79 (0.02) | 0.28 (0.03) | 0.70 (0.09) | 0.75 (0.03) |
| middle | 0.51 (0.05) | 0.25 (0.06) | 0.65 (0.11) | 0.55 (0.03) |
| low | 0.21 (0.02) | 0.3 (0.11) | 0.64 (0.9) | 0.25 (0.03) |

*task-specific information is presented here as the average mutual information (MI) and the standard deviation of each fragment set relative to the mutual information of a task-ideal fragment (one that is detected in all and only those instances in which the class of interest is present).

doi:10.1371/journal.pone.0003978.t001

series there were four stimulus blocks separated by 30-second fixation blocks during which participants were instructed to look at a fixation cross displayed in the center of the screen. In total, we acquired 6 time series with image fragments and 2 additional time series with blocks of faces and objects serving as a standard face-localizer.

On every trial the participants performed an unrelated task. The stimuli were randomly jittered 1° to the left/right of the center fixation cross and the participants performed a one-back location task by pushing one of two buttons randomly assigned with the index fingers of the two hands.

**Scanning parameters.** Scanning was carried out at the Brown University MRI Research Facility with a Siemens 3T TIM Trio magnet with an 8-channel phased-array head coil. Functional images were acquired with an ascending interleaved echo-planar imaging (EPI) pulse sequence (90 time points per time series; TR = 3 s; TE = 30 ms; flip angle 90°; 3 mm isotropic voxels; field of view $192\times192\times144$ $mm^3$; 48 slices covering the entire cerebral cortex). At the beginning of each session, we also acquired a T1-weighted anatomical image (1 mm isotropic voxels; 160 slices of size $256\times256$ mm).

**Analysis of imaging data.** Analysis was carried out using AFNI [33] and custom in-house Matlab (Mathworks, Natick MA) code. The first 5 images of each fMRI time series, during which subjects maintained fixation, were removed to allow the hemodynamics to achieve a steady state and to minimize transient effects of magnetic saturation. Further preprocessing involved slice scan time correction, 3-D motion correction, smoothing with a Gaussian kernel of 6 mm FWHM, normalization (each voxels's time series was divided by its mean intensity to convert the data from arbitrary image intensity units to percent signal modulation) and linear trend removal. Group analyses were performed after converting functional images into Talairach space [34].

Conventional univariate mapping analysis was performed on each participant's data. For each experimental condition we constructed a box-car predictor and convolved it with a gamma function. The general linear model [35] was applied to compute the coefficient of each predictor independently for each voxel. Significance maps of the brain were computed by t-tests of pairwise comparisons between relevant conditions. Significance levels were corrected by taking into account cluster size and its false detection probability[36] ($p = 0.05$ corrected). This type of analysis was used to contrast the high and low information conditions for detection and individuation as well as faces versus objects in a standard face localizer test [7,37].

In addition, we performed multivariate pattern analysis [38] to distinguish between the two levels of task-specific information for each task in face selective areas. Principal component analysis (PCA) was first applied to the coefficients of all voxels in a given area across all blocks to reduce the dimensionality of the patterns. A 'leave-one-out' ('jackknife') classification procedure was then carried out on the resultant patterns. More precisely, we trained a linear classifier on the patterns corresponding to all blocks except one and tested it on this remaining pattern. This procedure was repeated in turn for all blocks, every time leaving out a different pattern. For the purposes of pattern classification we used a linear support vector machine (SVM)–other classifiers we tested, such as a single-layer perceptron, yielded similar results. Importantly, multivariate analysis was carried out on a version of the data that had not been spatially smoothed, thus preserving high-frequency spatial information.

## Results

### Task-Specific Fragment Information for Detection and Individuation

Task-specific information values for fragment types reliably transferred from the training face dataset to a new dataset. The correlation between the scores for the training and the test dataset were significant for both detection ($r = 0.93$, $p<0.001$) and individuation ($r = 0.92$, $p<0.001$). Fragments at three levels of information, high, middle and low,–see Table 1–are shown in Figure 2 for detection and in Figure 3 for individuation. While fragments at each level of information covered together most of the face, we note several tendencies. In the case of detection, the most diagnostic fragments tended to span the area between the eyes, less diagnostic ones tended to contain only one feature such as one eye or the nose and the least diagnostic ones contained the hairline or the chin. For individuation, highly and intermediately diagnostic fragments contained the top part of the face while the least informative ones contained the lower part of the face, the chin, the mouth and the lower nose.

As far as the relationship between the two types of task-specific information is concerned, the comparison of the scores for detection and individuation showed a weak albeit significant correlation both within the training set ($r = 0.25$, $p<0.001$) and within the test set ($r = 0.23$, $p<0.001$). These results suggest that the two types of information may be roughly separable from one another. In line with this suggestion and providing further confirmation for it, we were able to manipulate one type of task-specific information independently of the other when selecting our experimental stimuli while controlling at the same time for low-level properties of the fragments.

### Behavioral Results–Experiments 1 and 2

None of the participants were able to correctly identify any individuals familiar to them from experience prior to the experiments.

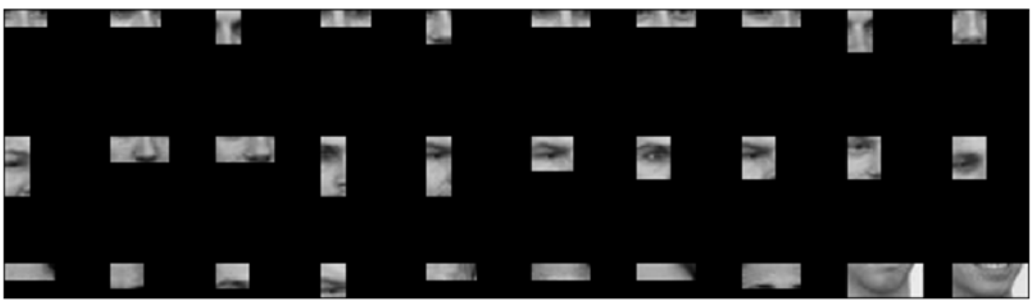

**Figure 2. Face fragments with high (top), intermediate (middle) or low (bottom) levels of detection-specific information.**
doi:10.1371/journal.pone.0003978.g002

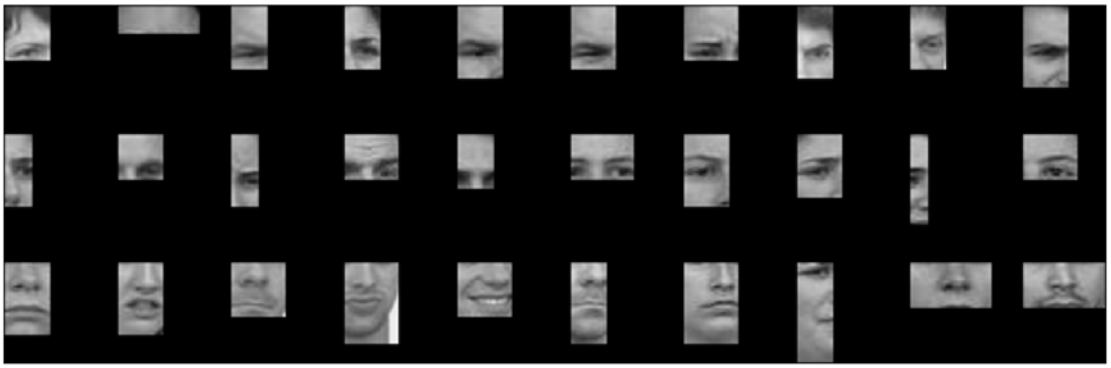

**Figure 3. Face fragments with high (top), intermediate (middle) or low (bottom) individuation-specific information.**
doi:10.1371/journal.pone.0003978.g003

Accuracy scores for each participant were computed using d', a signal detection measure of discrimination performance between two classes combining the relative contribution of hits and false alarms [39]. In the case of detection, hits and false alarms were provided by correct and incorrect 'face' responses, while for individuation they were provided by correct and incorrect 'same-individual' responses.

Repeated measures analysis of variance was conducted across the discrimination performance of the participants in each experiment. We found a main effect of information level for both detection ($F(15, 30) = 3.92$, $p<0.001$) and individuation ($F(15, 30) = 4.88$, $p<0.001$) indicating that participants are less accurate at recognizing faces with decreasing amounts of relevant information–Figure 4. In addition, we performed a two-way analysis of variance across the level of information and task combining the results of the two experiments. The analysis revealed significant effects for both the task factor ($F(1, 90) = 231.71$, $p<0.001$) and the interaction of the two factors ($F(2, 90) = 7.33$, $p<0.01$). We also note that performance was above chance for all information levels in both experiments (significantly above d' = 0).

Similar analyses computed for reaction times revealed no significant effect of task-specific information for either task ($p>0.1$)–Figure 5. In the two-way analysis of variance, we found a main effect of task ($F(1, 90) = 23.63$, $p<0.001$) but no significant interaction ($p>0.5$). The main effect of the task is a good indicator of task difficulty: face individuation was more difficult than face detection despite the larger size of the stimuli as reflected by both discrimination performance and reaction times.

## fMRI Results

Group maps of task-specific information effects were obtained by averaging the statistical parametric maps of individual participants–Figure 6. The comparison of the two detection conditions revealed two areas more active for diagnostic fragments than non-diagnostic fragments across participants ($p<0.05$,

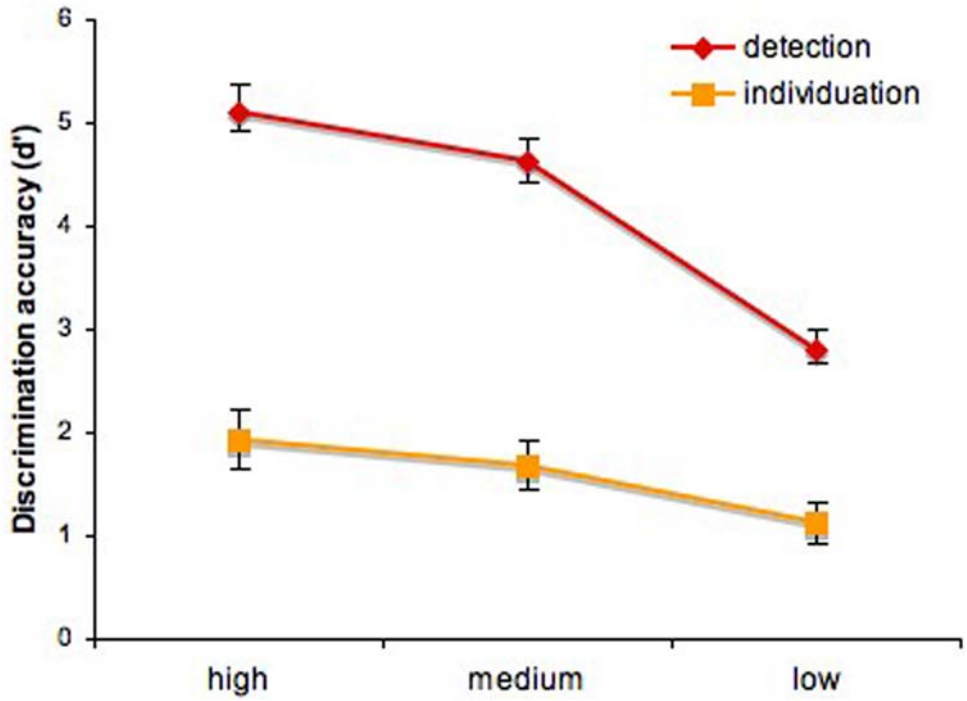


**Figure 4. Discrimination accuracy for detection and individuation across three levels of task-specific information (mean±SEM).**
doi:10.1371/journal.pone.0003978.g004

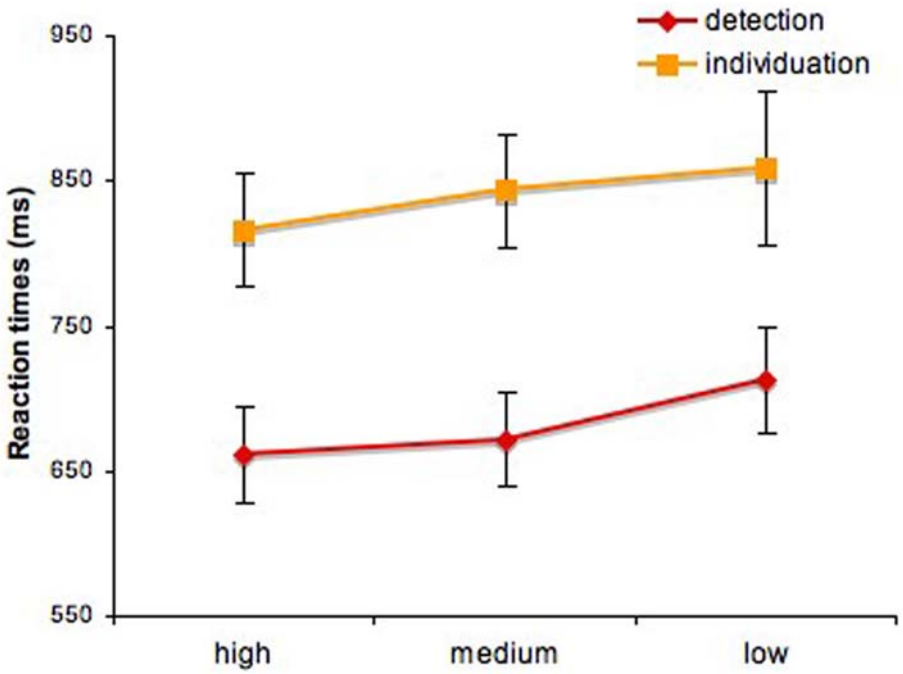


**Figure 5. Reaction times for detection and individuation across three levels of task-specific information (mean±SEM).**
doi:10.1371/journal.pone.0003978.g005

corrected): a region in the right posterior fusiform gyrus (pFG) and a bilateral region in the inferior occipital gyrus (IOG)–see Table 2. The first of these regions borders the functionally-localized face-selective region we find in the right fusiform gyrus, also known as the right 'fusiform face area' (FFA) [27], while the second surrounds and completely contains the functionally-localized face-selective region we found in the right IOG, the right 'occipital face area' (OFA) [26]. (Consistent with other studies, the left OFA was not reliably found across a number of participants and therefore excluded from analyses). In contrast, the comparison of the two individuation conditions revealed one area more active in the right anterior inferior temporal gyrus (aIT).

Specific region-of-interest analyses were performed in the functionally-localized face-selective areas nominally forming the core system for face processing [3]: the FFA, the OFA and another region located bilaterally in the superior temporal sulcus (STS). The areas were individually localized across participants using the data from the standard face-localizer scans, and the average percentage signal change was computed for each area and each task. Detection effects were reliably found in the right FFA ($t(7) = 4.14$, $p<0.05$) and the right OFA ($t(7) = 3.22$, $p<0.05$)–Figure 7. In contrast, no individuation-specific information effects were found in any of these regions ($p>0.1$).

In addition to the analysis of face-selective regions of interest, we examined the response of regions localized for one task, with the other task as well; for example, we examined whether there is sensitivity to detection-specific information in the aIT region already identified as sensitive to individuation-specific information. No significant effects were found for any of these comparisons.

Next, pattern analysis was applied across blocks within each face-selective region for each subject. The discriminability of the two levels of information for each task was encoded using again the d' measure. Neural responses elicited by higher-information fragments were encoded as hits or false alarms when recognized correctly and incorrectly, respectively. The only region that showed significant sensitivity across participants was the right FFA when viewing fragments varying across levels of individuation-specific information ($t(7) = 3.67$, $p<0.01$)–Figure 8. No other region was significantly different from chance for either task ($p>0.05$).

## Discussion

Our study examines how different tasks impact the featural code used in human face processing. From the many tasks that constitute face recognition, we focused on detection and individuation in that they represent two ends of the face recognition spectrum. This comparison is made tractable by adopting a general computational framework–fragment-based category representations. The concrete question we ask within this framework is twofold: how does the task-specific information carried by face fragments objectively vary within and across tasks and how sensitive is the visual system to these types of variation?

First, from a computational perspective, we find that the two types of task-specific information, for detection and individuation, are roughly separable when considering the mutual information between fragment presence and the category of interest. This result is not entirely unexpected given that in order to be diagnostic for the two tasks a fragment type would need to satisfy two criteria presumably at tension with each other. For detection, the area of the face captured by a fragment would need to exhibit small *extrapersonal* variability and be visually dissimilar from recurrent non-face image structures. For individuation, on the other hand, the same area would need to exhibit large *extrapersonal* variability relative to *intrapersonal* variability [30]. Consistent with this tension, the correlation we found between the two types of task-specific information is relatively small although still significant. The small size of this correlation is what justifies and explains our ability to select (with relative ease) two subsets of fragments that vary independently with respect to their task-specific information for the two tasks. We then use these separate subsets to examine the relationship between detection and identification in

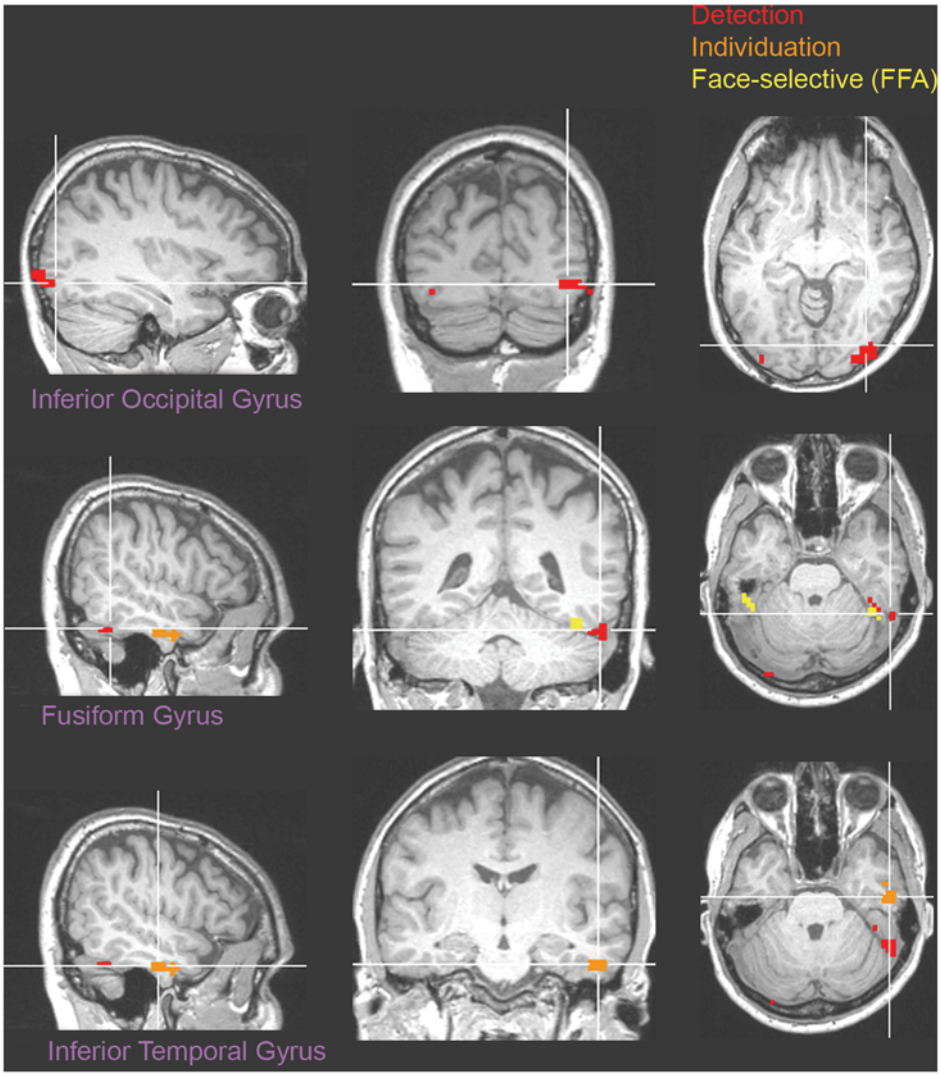


**Figure 6. Group map superimposed on the brain of one participant.**
doi:10.1371/journal.pone.0003978.g006

human visual processing–a high correlation would have made this analysis less likely to succeed. To be clear, we are not using this small, but significant correlation to argue alone for the separability of detection and individuation features, nor are we arguing for complete separability. Instead we suggest that partial but reliable separability occurs with regard to task-specific features. Based on these results, it would appear, detection-diagnostic fragments should still be able to support individuation, albeit in a non-optimal fashion, and vice versa. The extent to which this prediction holds for automatic recognition should be the subject of further investigation. Interestingly, our neuroimaging results hint that this may indeed be the case with face processing in the human visual system.

**Table 2.** Areas sensitive to task-specific information.

| Task | Region | Coordinates (center) | | | Size ($mm^3$) | Peak *t*-value |
|---|---|---|---|---|---|---|
| | | x | y | z | | |
| detection | R.IOG | 33 | −86 | −6 | 1647 | 6.98 |
| detection | L.IOG | −36 | −83 | −10 | 648 | 5.64 |
| detection | R.FG | 48 | −45 | −25 | 621 | 4.50 |
| individuation | R.aIT | 50 | −9 | −28 | 702 | 4.33 |

doi:10.1371/journal.pone.0003978.t002

Second, our behavioral and neuroimaging results indicate that the human visual system is independently sensitive to information diagnostic for both detection and individuation. Behaviorally, visual recognition performance with image fragments improves with increasing amounts of task-specific information carried by face fragments for both tasks. With respect to neuroimaging, a number of regions in the ventral visual pathway were found that respond more robustly to fragments carrying higher levels of task-specific information relative to fragments carrying lower levels of information for both tasks. Region-of-interest analyses revealed functionally-defined face-selective regions, such as the right FFA and OFA, also showed sensitivity to detection-specific information. Given that the procedure used to localize these regions is a form of face detection, that is, comparing faces to objects, this may not be very surprising. However, we note that detection sensitivity was tested

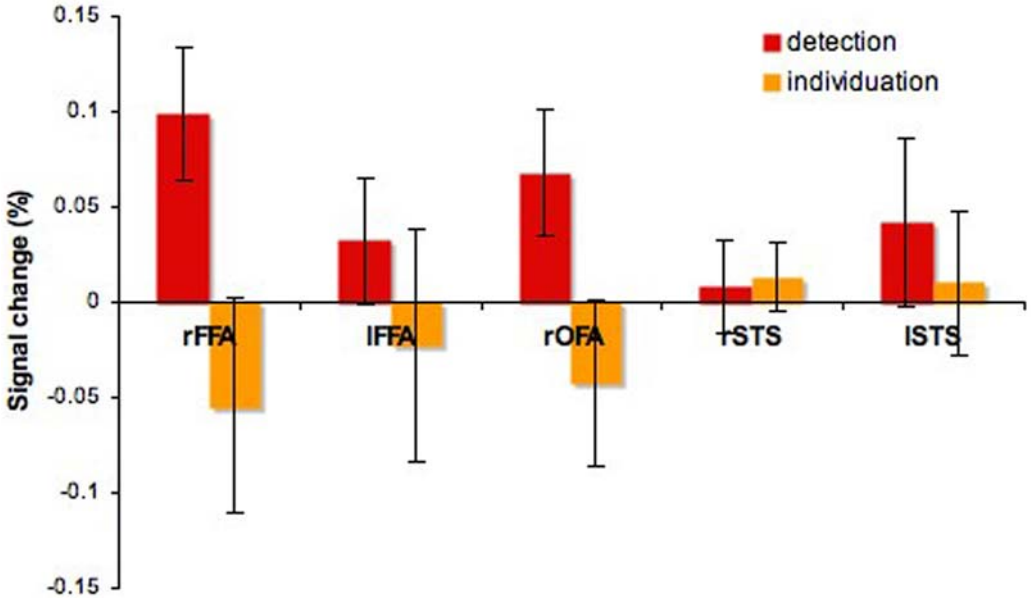


**Figure 7. Task-specific information effects in face-selective regions (mean±SEM).**
doi:10.1371/journal.pone.0003978.g007

using *only* face fragments and these fragments covered less than a fifth of the area of a whole face. This suggests that some face parts are preferentially represented relative to others given their informativeness with regard to face detection. That the neural coding of features is sensitive to the demands of face detection has been previously noted [14] and is consistent with our present results. Here we show that such results, presumably due to detection sensitivity, are independent from individuation. Moreover, we extend such results to individuation and we conclude that the visual system responds to the constraints imposed by both tasks.

Third, the neural representation of faces appears to differentially reflect detection and individuation demands. Sensitivity to the former is revealed by the size of the neural response in a series of regions in the bilateral IOG and the right pFG while sensitivity to the latter is found both in the size of the neural response in one area of the right aIT as well as in the neural pattern in the right FFA. Overall, these findings support the idea of different types of neural representations underlying detection or individuation.

One specific point of contention regards the role of the FFA and the aIT in these two tasks. A considerable body of results from neuroimaging [7,6,40,5,4,41] and neuropsychology [42–44] suggests that the FFA is involved in detection and individuation. However, at least one study [8] challenges this view and suggests the FFA may delegate other regions, in particular the aIT, to process faces at the individual level. On this account, the sensitivity of the FFA to face individuation demands, as revealed by neuroimaging results, could simply be due to the feedback received from such regions rather than because of its direct involvement in the task.

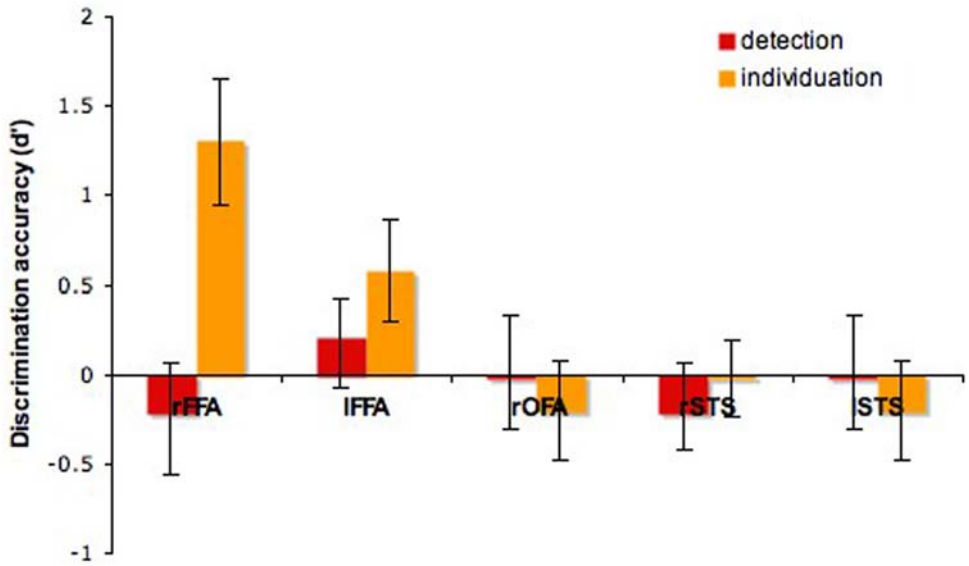


**Figure 8. Pattern discrimination performance for the two tasks in face-selective regions (mean±SEM).**
doi:10.1371/journal.pone.0003978.g008

An interesting perspective on this issue comes from the study of congenital prosopagnosia, a condition characterized by profound impairment in face recognition, particularly at the individual level, in the absence of an obvious insult to the brain. A recent diffusion tensor imaging study [10] associated recognition performance in prosopagnosics with the degree of structural integrity of the right Inferior Longitudinal Fasciculus (ILF). ILF is one of the two major fiber tracts passing through the fusiform gyrus and connects the lingual and fusiform gyri with the superior, inferior and middle temporal gyri as well as the hippocampus and the parahippocampus. While, given the length of the tract, this result by itself may fail to directly involve the aIT, it could explain cortical volume alterations in the inferotemporal cortex observed in this population [45]. Interestingly, activation in face-selective areas in prosopagnosics and normal humans appears to be comparable [46,47]–but see [44]. This pattern of results seems to suggest the aIT is important for face recognition and a partial breakdown in the communication between the fusiform gyrus and the aIT may be a plausible source of face individuation deficits.

Our results can help bridge two potentially divergent lines of evidence. In agreement with most neuroimaging studies, we find evidence for the direct involvement of the FFA in face individuation and against the hypothesis of an indirect feedback-conditioned role. At the same time, we also find evidence for the role of the aIT in individuation [8], a role also suggested by the neuropsychological literature. However, the FFA and the aIT turned out to exhibit two different types of sensitivity to individuation, one revealed by multivariate pattern analysis and the other by univariate analysis. This difference by itself does not explain, of course, why the same studies fail to involve both areas in individuation–they typically implicate only the FFA or aIT, but not both. In addition to the difference in sensitivity revealed by our two analysis methods, this discrepancy can be accounted for in several other ways. First, face-localizer tests are optimized primarily for detection, that is, they compare faces to other categories of objects, and thus can fail to involve neural structures that serve primarily face individuation. Second, given the special status conferred to a group of regions including the FFA as the 'core system' for face processing [3], neuroimaging research has particularly focused on these restricted brain areas and, thus, may fail to observe relevant activation in other areas, for example, the aIT. Third, and most importantly, our stimuli were face fragments selected based on their task-specific information instead of whole faces. Thus, our study is aimed at dealing in a more direct manner with individuation sensitivity than previous studies using whole face images.

Overall, we interpret our current results as supporting the involvement of the FFA primarily in detection and of the right aIT in individuation. However, individual face differences are already represented in the FFA. These differences seem to be further amplified and recoded in the right aIT insofar as they lead to different types of sensitivity to individuation-specific information. If this is the case, we expect the FFA to support individuation without the help of the aIT at least to some extent. However, if the features encoded in the FFA serve primarily face detection, individuation processing in this area is likely to be suboptimal. This motivates and explains the recruitment of a different area, the aIT, dedicated to a task critical in our everyday life, individuation.

Our current results are also interesting from the perspective of the hierarchy of visual processing along the ventral visual pathway [48]. The idea that visual features of increasing complexity build successively upon each other at different levels of visual processing has been incorporated in many neurally-inspired models of object [16,49] and face recognition [17]. More recently, this approach has also been extended to fragment-based processing [50,51]. Composing larger, more specialized fragments successively out of smaller and more generic fragments across a series of representational levels is a computationally attractive means for instantiating hierarchical processing. Relevant to our study, this hints that larger individuation-dedicated fragments may be separately represented and built upon smaller detection-dedicated fragments. One concrete possibility suggested by the present results is that features optimal for individuation are represented as *patterns* over face detection features within the FFA and then recoded in a more localist fashion within the right aIT. The neural plausibility of this hypothesis is the goal of further research.

Finally, our results reinforce the assumption that overlapping image fragments provide a neurally-plausible representational schema for object features. The argument the present study makes for their plausibility and utility is that they help clarify how different tasks shape the neural code underlying face recognition. However, we should also acknowledge the limits of this approach as an actual model. One question regards the effectiveness of using rectangular fragments to represent what are, most likely, non-rectangular features. In response to this, we take rectangular-shaped features to be a rough but reasonable approximation of the actual features encoded by the visual system. More plausible features with smoother edges without sharp corners should be furthered examined. However, we argue, our investigation of task-specific information as carried by fragments is systematic and sufficiently detailed that the precise shape of the features should not alter significantly the conclusions we reached above. Consistent with this, replacement of square-like features with circular ones did not significantly alter measurements of detection-specific information in a related study [14]. Another more critical issue concerns the manner in which fragments are actually represented by the neural code. For instance, every fragment contains a multitude of cues with different contributions to various recognition tasks. How such cues are separately considered, encoded and integrated into a unified representation is an important problem that should be addressed further. For present purposes, we treat rectangular image patches as reasonable stand-ins for the actual biological representations of different face parts and subparts.

To conclude, we examined and found that face detection and individuation place different task constraints on the representational code required for automatic and human face recognition. More generally, we interpret these results as further evidence for the soundness of fragment-based models of human object processing.

## Supporting Information

**Figure S1** Training set of face images
Found at: doi:10.1371/journal.pone.0003978.s001 (0.48 MB TIF)

**Figure S2** Natural image fragments erroneously labeled as face fragments by the method
Found at: doi:10.1371/journal.pone.0003978.s002 (0.08 MB TIF)

## Acknowledgments

We thank the members of the Perceptual Expertise Network (PEN) for many helpful and insightful comments as well as the staff of the Brown University Magnetic Resonance Facility for assistance with data collection.

## Author Contributions

Conceived and designed the experiments: AN JMV MJT. Performed the experiments: AN JMV. Analyzed the data: AN JMV. Wrote the paper: AN JMV MJT.

## References


1. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. ACM Computing Surveys 35: 399–458.
2. Bruce V, Young A (1986) Understanding Face Recognition. British Journal of Psychology 77: 305–327.
3. Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4: 223–233.
4. Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat Neurosci 8: 107–113.
5. Pourtois G, Schwartz S, Seghier ML, Lazeyras F, Vuilleumier P (2005) Portraits or people? Distinct representations of face identity in the human visual cortex. J Cogn Neurosci 17: 1043–1057.
6. Grill-Spector K, Knouf N, Kanwisher N (2004) The fusiform face area subserves face perception, not generic within-category identification. Nat Neurosci 7: 555–562.
7. Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JC, Anderson AW (2000) The fusiform "face area" is part of a network that processes faces at the individual level. J Cogn Neurosci 12: 495–504.
8. Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A 104: 20600–20605.
9. Thomas C, Moya L, Avidan G, Humphreys K, Jung KJ, et al. (2008) Reduction in white matter connectivity, revealed by diffusion tensor imaging, may account for age-related changes in face perception. J Cogn Neurosci 20: 268–284.
10. Thomas C, Avidan G, Humphreys K, Jung K, Gao F, Behrmann M (in press) Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nat Neurosci.
11. Liu J, Harris A, Kanwisher N (2002) Stages of processing in face perception: an MEG study. Nat Neurosci 5: 910–916.
12. Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5: 682–687.
13. Harel A, Ullman S, Epshtein B, Bentin S (2007) Mutual information of image fragments predicts categorization in humans: Electrophysiological and behavioral evidence. Vision Res 47: 2010–2020.
14. Lerner Y, Epshtein B, Ullman S, Malach R (2008) Class information predicts activation by object fragments in human object areas. J Cogn Neurosci 20: 1189–1206.
15. Hegdé J, Bart E, Kersten D (2008) Fragment-based learning of visual object categories. Curr Biol 18: 597–601.
16. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2: 1019–1025.
17. Jiang X, Rosen E, Zeffiro T, VanMeter J, Blanz V, Riesenhuber M (2006) Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. Neuron 50: 159–172.
18. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3: 71–86.
19. Hancock PJB, Burton AM, Bruce V (1996) Face processing: Human perception and principal components analysis. Memory and Cognition 24: 26–40.
20. Nestor A, Tarr MJ (2008) The segmental structure of faces and its use in gender recognition. J Vis 8: 1–12.
21. Balas BJ, Sinha P (2006) Region-based representations for face recognition. ACM Transactions on Applied Perception 3: 354–375.
22. Maurer D, Grand RL, Mondloch CJ (2002) The many faces of configural processing. Trends Cogn Sci 6: 255–260.
23. Zhang L, Cottrell GW (2005) Holistic processing develops because it is good. Proceedings of the Cognitive Science Society. pp 2428–2433.
24. Neri P, Levi DM (2006) Receptive versus perceptive fields from the reverse-correlation viewpoint. Vision Res 46: 2465–2474.
25. Murray RF, Bennett PJ, Sekuler AB (2002) Optimal methods for calculating classification images: Weighted sums. J Vis 2: 79–104.
26. Eckstein MP, Ahumada AJ (2002) Classification images: a tool to analyze visual strategies. J Vis 2: 1x.
27. Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res 41: 2261–2271.
28. Gosselin F, Schyns PG (2004) No troubles with bubbles: a reply to Murray and Gold. Vision Res 44: 471–477.
29. Cover T, Thomas J (1991) Elements of information theory. New York: Wiley.
30. Moghaddam B, Jebara T, Pentland A (2000) Bayesian face recognition. Pattern Recognition 33: 1771–1782.
31. Brainard DH (1997) The psychophysics toolbox. Spat Vis 10: 433–436.
32. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat Vis 10: 437–442.
33. Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29: 162–173.
34. Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system - an approach to cerebral imaging. New York: Thieme Medical Publishers.
35. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1995) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2: 189–210.
36. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33: 636–647.
37. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17: 4302–4311.
38. O'Toole AJ, Jiang F, Abdi H, Penard N, Dunlop JP, Parent MA (2007) Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J Cogn Neurosci 19: 1735–1752.
39. Snodgrass JG, Corwin J (1988) Pragmatics of measuring recognition memory: applications to dementia and amnesia. J Exp Psychol Gen 117: 34–50.
40. Loffler G, Yourganov G, Wilkinson F, Wilson HR (2005) fMRI evidence for the neural representation of faces. Nat Neurosci 8: 1386–1390.
41. Winston JS, Henson RN, Fine-Goulden MR, Dolan RJ (2004) fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J Neurophysiol 92: 1830–1839.
42. Barton JJ, Press DZ, Keenan JP, O'Connor M (2002) Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. Neurology 58: 71–78.
43. Damasio AR, Damasio H, Van Hoesen GW (1982) Prosopagnosia: anatomic basis and behavioral mechanisms. Neurology 32: 331–341.
44. Hadjikhani N, de Gelder B (2002) Neural basis of prosopagnosia: an fMRI study. Hum Brain Mapp 16: 176–182.
45. Behrmann M, Avidan G, Gao F, Black S (2007) Structural imaging reveals anatomical alterations in inferotemporal cortex in congenital prosopagnosia. Cereb Cortex 17: 2354–2363.
46. Hasson U, Avidan G, Deouell LY, Bentin S, Malach R (2003) Face-selective activation in a congenital prosopagnosic subject. J Cogn Neurosci 15: 419–431.
47. Avidan G, Hasson U, Malach R, Behrmann M (2005) Detailed exploration of face-related processing in congenital prosopagnosia: 2. Functional neuroimaging findings. J Cogn Neurosci 17: 1150–1167.
48. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1: 1–47.
49. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A 104: 6424–6429.
50. Ullman S (2007) Object recognition and segmentation by a fragment-based hierarchy. Trends Cogn Sci 11: 58–64.
51. Epshtein B, Lifshitz I, Ullman S (2008) Image interpretation by a single bottom-up top-down cycle. Proc Natl Acad Sci U S A 105: 14298–14303.

Human Brain Mapping

# Internal representations for face detection – an application of noise-based image classification to BOLD responses

| | |
|---|---|
| Journal: | *Human Brain Mapping* |
| Manuscript ID: | HBM-11-0937.R2 |
| Wiley - Manuscript type: | Research Article |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Nestor, Adrian; Carnegie Mellon University, Center for the Neural Basis of Cognition<br>Vettel, Jean; US Army Research Laboratory,<br>Tarr, Michael; Carnegie Mellon University, Center for the Neural Basis of Cognition |
| Keywords: | face recognition, reverse correlation, fMRI |
| | |

SCHOLARONE™
Manuscripts

For Review

# Internal representations for face detection – an application of noise-based image classification to BOLD responses

Short title: Internal representations for face detection

Adrian Nestor [ab]*

Jean M. Vettel [c]

Michael J. Tarr [ab]

[a] Center for the Neural Basis of Cognition, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213

[b] Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213

[c] US Army Research Laboratory, Aberdeen Proving Ground, MD

*To whom correspondence should be addressed.

Phone: (412) 268-4237

Fax: (412) 268-2798

E-mail: anestor@andrew.cmu.edu

Keywords: face recognition, reverse correlation, fMRI

For Peer Review

**Abstract**. What basic visual structures underlie human face detection and how can we extract such structures directly from the amplitude of neural responses elicited by face processing? Here, we address these issues by investigating an extension of noise-based image classification to BOLD responses recorded in high-level visual areas. First, we assess the applicability of this classification method to such data and, second, we explore its results in connection with the neural processing of faces. To this end, we construct luminance templates from white noise fields based on the response of face-selective areas in the human ventral cortex. Using behaviorally and neurally-derived classification images, our results reveal a family of simple but robust image structures subserving face representation and detection. Thus, we confirm the role played by classical face selective regions in face detection and we help clarify the representational basis of this perceptual function. From a theory standpoint, our findings support the idea of simple but highly diagnostic neurally-coded features for face detection. At the same time, from a methodological perspective, our work demonstrates the ability of noise-based image classification in conjunction with fMRI to help uncover the structure of high-level perceptual representations.

For Peer Review



John Wiley & Sons, Inc.

## Introduction

Extensive research has focused on mapping out the neural resources involved in face processing [Gauthier et al., 2000; Haxby et al., 2000; Ishai et al., 2005; Kanwisher et al., 1997; Rossion et al., 2003] and on exploring how these resources enable various recognition tasks such as detection or individuation (Fox et al., 2009; Kriegeskorte et al., 2007; Nestor et al., 2011; Pourtois et al., 2005; Winston et al., 2004). However, our understanding of the visual representations underlying recognition is far more limited. An obvious example in this sense is face detection: while the face selectivity of certain cortical areas, such as the fusiform face area (FFA), has been commonly associated with detection [Avidan et al., 2005; Freiwald et al., 2009; Loffler et al., 2005; Tong et al., 2000], it is still unclear how these areas are able to perform this function. Theoretically, uncovering the representational basis of face detection is critical in that detection precedes and constrains other face processing tasks [Liu et al., 2002; Or and Wilson, 2010; Tsao and Livingstone, 2008]. Methodologically though, uncovering the structure of neural representations poses a significant challenge.

A standard approach to exploring internal representations for visual recognition involves hypothesis-testing. Specifically, one can select a biologically-plausible recognition schema, adopt it as a model of neural processing and test its validity. For instance, a detection schema relying on image fragments [Ullman et al., 2002] has been tested with relative success as a model of neural face processing [Harel et al., 2007; Nestor et al., 2008]. However, this approach is limited by the specificity of the representational types assumed (e.g., do neural representations encode actual image fragments?) and by the difficulty of their interpretation (e.g., what properties of a fragment underlie its diagnosticity for face detection?).


John Wiley & Sons, Inc.
For Peer Review

Another approach, more challenging but less restrictive in terms of theoretical assumptions, involves reconstructing the relevant visual features rather than testing a specific class. Reverse correlation methods have been extensively employed to this effect in neurophysiology and behavioral research [Neri and Levi, 2006; Ringach and Shapley, 2004]. A family of such techniques, known as image classification [Abbey and Eckstein, 2002; Ahumada, 2002; Beard and Ahumada, 1998; Gold et al., 2000], achieve this goal by combining noise fields into a unique template based on the discrete responses they elicit. This template, referred to as a 'classification image' (CI), serves as an approximation of the image structure that accounts best for a given set of responses. As the elements entering the construction of a CI are typically structure-free (e.g., white noise fields), it is inferred that any significant structure apparent in the CI lies with the source of the responses, be that a single neuron or a behavioral subject. However, this approach is costly in terms of the number of trials needed and restrictive with respect to the type of features targeted, i.e., prominent, robust, simple features.

Here, we derive face detection templates by applying image classification to behavioral and neural responses recorded in the human ventral cortex. To deal with the challenge of applying noise-based image classification to BOLD data, we consider several ways of optimizing the quality of our neurally-derived CIs. First, we collected a relatively large number of trials by testing each subject across multiple scanning sessions (12-13). Second, we took advantage of the continuous nature of the BOLD signal by adapting a suitable version of image classification [Murray et al., 2002] to our data (i.e., a version not restricted to binary responses). Third, we used slow event trials allowing us to maximize the SNR of trial-specific BOLD responses. Finally, we constructed CIs corresponding to face selective regions as established by an independent "localizer". This fact is important in that standard image classification is based on a


John Wiley & Sons, Inc.
For Peer Review

linearity assumption: the magnitude / likelihood of a response increases linearly with the presence of a particular image structure. Recent research shows that face-selective areas, but not other high-level visual areas, exhibit this property in response to faces [Davidenko et al., 2011; Horner and Andrews, 2009] warranting the application of image classification to their responses.

Of note, face detection is a suitable domain for assessing the application of image classification to BOLD data. Faces, as a visual category, are remarkably homogeneous and, thus, likely to contain a few highly diagnostic detection features [Sinha, 2002]. The presence of such features and the robustness of their encoding are key factors in being able to derive meaningful visual features from BOLD data. Supporting this idea, a number of recent electroencephalography (EEG) and behavioral results [Hansen et al., 2010; Rieth et al., 2011; Smith et al., 2012] suggest that image classification is a viable and promising approach to the study of face perception (see Discussion).

As far as the regions targeted by our investigation are concerned, the FFA naturally holds particular interest. This is due not only to the prominent role played by the FFA in face perception [Kanwisher and Yovel, 2006] but also to its sensitivity to unconsciously processed face stimuli [Jiang and He, 2006] and even to stimuli erroneously expected to contain faces [Righart et al., 2010; Zhang et al., 2008]. These latter results are particularly relevant in that the ability of noise-only trials to elicit activation in the FFA is critical for the application of image classification to FFA responses. At the same time, we note that it is important to extend such investigations, insofar that it is possible, beyond the FFA to other high-level visual areas (and to other face-selective regions in particular).

In short, our work explores the representational basis of face detection associated with neural face processing. In this context, the use of noise-based image classification serves a

For Peer Review



John Wiley & Sons, Inc.

twofold purpose by allowing us to uncover the basic image structures underlying face detection and, more generally, by providing the opportunity to assess the applicability of this method to fMRI.

## Methods

### *Subjects*

Two young adults, EC and EA (both female, 22 years old), volunteered to participate in the experiment in exchange for payment. Subjects were right-handed, with normal (or corrected-to-normal) vision and no history of neurological disorder. Both subjects provided written consents. All procedures were approved by the Institutional Review Board of Brown University.

### *Stimuli*

An average face base image was constructed by combining multiple frontal-view faces from the Max Plank Institute, Tübingen (MPIK) face dataset (the current version is available at http://faces.kyb.tuebingen.mpg.de) – see Fig. 1. The database contains 200 Caucasian faces of different individuals with neutral expressions collected under consistent lighting conditions. We converted all faces to grayscale, cropped them and normalized them with the position of the eyes and the nose. The base image was obtained by averaging individual faces along with their mirror symmetric versions and subsampling the resulting image to 38 X 32 pixels. The contrast of the base image was separately adjusted for each subject as described below (see Experimental paradigm).

For Peer Review



John Wiley & Sons, Inc.

Experimental stimuli were constructed half of the time by adding white Gaussian noise to the base image and half of the time from noise only (see Fig. 1 for examples). Noise had a fixed root-mean-square (RMS) Weber contrast of 27%. The size of the contrast was maximized under the constraint that all pixel luminance values within two standard deviations of the mean fall within a displayable range. Values outside this interval were discarded and resampled on each trial. The stimuli subtended a visual angle of 4.4° X 5.2° after tripling the size of the images by pixel replication.

We note that the effective resolution of our stimuli is relatively low allowing us to minimize the size of the search space. While this restricts the use of high-level frequencies in performing the task, it is unlikely to affect face detection in a critical manner given that the optimal band for face recognition in humans lies under 16 cycles per face width [Costen et al., 1996; Näsänen, 1999; Peli et al., 1994].

Stimulus design and presentation relied on Matlab 7.5 (Mathworks, Natick, MA) and the Psychophysics Toolbox 3 [Brainard, 1997; Pelli, 1997] running on an OS X Apple Macintosh.

*Imaging methods*

Each subject was scanned for 12 (EC) and 13 (EA) 1-hour sessions completed on separate days. Scanning was carried out using a Siemens 3T TIM Trio magnet with a 32-channel phased-array head coil. Functional images were acquired with an echo-planar imaging (EPI) pulse sequence (1.5 s TR; 36 ms TE; 90° flip angle; $2.2^3$ mm voxels; 193.6 X 193.6 X 39.6 mm FOV; 18 oblique slices covering the ventral stream). To maximize similarity in brain coverage across sessions, an anatomical landmark was selected for the top slice of the partial volume for

For Peer Review



John Wiley & Sons, Inc.

each subject. At the beginning of each session, we also acquired a T1-weighted anatomical image ($1^3$ mm voxels; 160 slices of total size 256 X 240 mm).

*Experimental paradigm*

Subjects performed a face detection task by discriminating noisy face stimuli from noise-only stimuli (Fig. 1). Both subjects were informed that half of the time the stimuli contained a face embedded in noise in the attempt to minimize bias. They were also informed that the face was the same on all trials, that it was of the same size as the rectangular stimulus appearing on each trial and that it was centrally located within the rectangle. Neither subject was exposed at any time to a noise-free version of the base image (i.e., they never saw the image shown in Fig. 1, left). Responses were made by pushing one of two buttons with the index fingers of both hands.

Each trial had the following structure: a high-contrast fixation cross was displayed for 100 ms followed by a stimulus for 400 ms followed, in turn, by a lower-contrast fixation cross for 10s. Thus, the duration of each trial totaled 10.5s.

During pilot testing, contrast thresholds for the base image corresponding to a 70% accuracy level were computed for each subject [Watson and Pelli, 1983] – 4.5% and 4.2% RMS Weber contrast for EC and EA, respectively. Noise contrast was the same across subjects and sessions. Each subject was tested across multiple days prior to scanning in order to ensure no further learning would take place (as reflected by better accuracy or shorter reaction times).

Each scanning session contained 5 - 7 face detection runs and each run consisted in 24 trials preceded by a 10.5 s fixation interval (for a total of 262.5s). Trial order was pseudo-randomized to maximize the uncertainty of stimulus category (noisy base image or just noise).


John Wiley & Sons, Inc.
For Peer Review

Across sessions we collected a total of 1920 and 2136 face-detection trials for EC and EA, respectively.

In addition, each session included one or two standard face-localizer runs for a total of 14 and 15 runs for EC and EA, respectively. During the localizer subject performed a category-unrelated task (monitoring for stimulus position) with faces, objects and scrambled images displayed in separate blocks**.** More specifically, each run contained 9 blocks (3 for each stimulus category), each block contained 15 trials and each trial consisted in 750ms of stimulus presentation followed by 250ms of fixation. Stimulus blocks were separated by 15s of fixation. Additional fixation blocks were introduced at the beginning and at the end of each run. The order of stimulus block types was counterbalanced across runs and no stimulus of any type was repeated within a session. Stimuli subtended a visual angle of approximately 3.9° X 4.3° and were randomly displayed on the left/right side of the fixation cross. On each trial subjects pressed one of two buttons associated with each position (left/right). The duration of each run totaled 285s.

In addition to identifying regions of interest (ROIs), the large number of localizers allowed us to verify the reproducibility of the ROIs and to obtain unbiased estimates of face selectivity in these regions [Kriegeskorte et al., 2010]. Critically, they also allowed us to monitor potential changes in the selectivity of these ROIs induced by a visually demanding task across numerous test sessions.

*Conventional analysis of imaging data*

Preprocessing steps involved slice scan time correction, 3-D motion correction, smoothing with a Gaussian kernel of 6 mm full-width half maximum (FWHM), normalization to



John Wiley & Sons, Inc.

For Peer Review

percent signal change and linear trend removal. All analyses were performed in the native space of each subject using AFNI [Cox, 1996] and custom Matlab code.

Face selective regions were localized through standard univariate analysis by contrasting blocks of faces and objects. Significance maps were corrected using the false discovery rate ($q < 0.05$). ROIs were further constrained by placing spherical masks (19 voxels) on the peak of each functionally defined area (Fig. 2).

In addition to the high-level visual areas mentioned above we also identified a control ROI in the calcarine sulcus of each participant. This early visual cortex (EVC) ROI was centered on the peak of another contrast, scrambled images versus objects and faces ($q < 0.05$), and was equated in terms of shape and size with the other ROIs. This particular region was chosen as a control ROI in order to assess the specificity of a number of effects to higher-level visual cortex (i.e., face-selective areas).

ROI mapping was performed using the first 5 localizer runs for each subject. Unbiased estimates of selectivity were computed using the remaining runs.

*CI computation*

Trials with no behavioral response and trials scoring reaction times significantly shorter / longer than the mean (±2SD) were discarded (5.6% and 8.8% for EC and EA). All analyses were performed on the remaining data.

Two classes of CIs were computed for each subject: behavioral CIs and neurally-derived CIs (based on BOLD data). Behavioral CIs were constructed by combining noise fields across trials following a standard approach [Ahumada, 2002; Beard and Ahumada, 1998]:


John Wiley & Sons, Inc.
For Peer Review

$$C = (\mu_{FF} + \mu_{NF}) - (\mu_{FN} + \mu_{NN}) \quad (1)$$

The terms $\mu_{FF}$ and $\mu_{NF}$ denote the average noise fields on trials on which subjects responded 'face' in the presence of a base image (hits) and in its absence (false alarms), respectively. Similarly, $\mu_{FN}$ and $\mu_{NN}$ denote the average noise fields on trials on which subjects responded 'noise' in the presence of a base image (misses) and in its absence (correct rejections). Figure 2 details the construction of the CIs and the outcome of this procedure for each subject.

The computation of neurally-derived CIs was performed as follows (see also Fig. 3). First, we computed average ROI amplitudes for each trial, normalized them (by z-scoring) and binned them separately by trial type (base image present / absent) and time point (1.5 through 10.5 seconds after stimulus onset). The size / number of bins is a parameter of the method – the results below were computed using a bin of size 0.4SD although smaller / larger bins produced similar results. Following this procedure every trial was labeled with its corresponding bin number (1 through 12 within the interval -2.4 to 2.4 SDs).

Second, for each ROI, time point -specific CIs were computed with the following formula [Murray et al., 2002]:

$$C = \sum_{i=1}^{n} \left(g(z_i) - g(z_{i+1})\right)\mu_{Fi} + \left(g(z_i - d') - g(z_{i+1} - d')\right)\mu_{Ni} \quad (2)$$

where $z_i = \frac{G^{-1}(p_{Fi-}) + G^{-1}(p_{Ni-}) + d'}{2}$.



John Wiley & Sons, Inc.

For Peer Review

Here $p_{Fi-}$ and $p_{Ni-}$ represent the probability of a bin number smaller than i when a base image is present and absent, respectively. $G^{-1}$ is the inverse of the normal cumulative distribution function, g is the normal probability density function, $d'$ is performance level and n represents the number of bins ($g(z_1)$ and $g(z_{n+1})$ are estimated as 0). Finally, $\mu_{Fi}$ and $\mu_{Ni}$ represent the average noise fields for a given bin i. The combination schema above represents an attempt to maximize the SNR of CIs derived from graded responses [Murray et al., 2002] and is here extended to work with BOLD responses.

Third, a single CI was computed for each ROI by taking a weighted sum over time-specific CIs using a standard hemodynamic response function [Friston et al., 1994]. This approach was expected to increase the overall SNR of the images (by computing a weighted average across multiple time points) and to produce a summary template of the visual structure driving the response of each ROI.

Finally, both behavioral and neural CIs were smoothed with a Gaussian filter with a 5-pixel FWHM allowing their analysis with random field theory-based tests [Chauvin et al., 2005].

## Results

*Face detection – behavioral performance and neural correlates*

Response accuracy across sessions was 69.3% and 73.2% for EC and EA. Subjects classified the stimuli as 'face' on 50.5% (EC) and 43.9% (EA) of the trials. Thus, as intended, both subjects reported the presence of a signal about half of the time (although we notice a small bias towards ‘no face’ responses in EA’s case).


John Wiley & Sons, Inc.
For Peer Review

The effect of trial type on neural responses was assessed for traditional ventral cortical regions for face processing [Haxby et al., 2000]. In both subjects, conventional functional contrasts (faces versus objects) revealed bilaterally the FFA [Kanwisher et al., 1997] and the occipital face area (OFA) [Gauthier et al., 2000] – Fig. 4 and Table 1. Average ROI responses (Fig. 5) were subjected to a three-way ANOVA (stimulus type x response type x ROI) across sessions separately for each subject. Here, stimulus type encodes whether the stimulus contained a base image or not while response type encodes the behavioral response, 'face' or 'no face'.

The analysis revealed significant effects for response type in each subject (EC: $F(1, 11)=27.25$, $p<0.001$; EA: $F(1, 12)=10.83$, $p<0.01$) and a significant interaction between response type and ROI for EA ($F(3, 36)=6.18$, $p<0.01$). To examine the source of this interaction we performed further contrasts that revealed significant effects of response type within each of EA's ROIs with the exception of the left FFA ($p>0.10$). At the same time, we did not find a main effect or interaction with stimulus type for either subject. Finally, a control area in the early visual cortex did not show any significant effects ($p>0.05$).

These results are important in several respects. First, they confirm the sensitivity of the FFA to face detection independent of the objective presence of a face stimulus [Righart et al., 2010; Zhang et al., 2008] and also extend this sensitivity to the OFA (as evidenced by response type effects). Second, they show that there is marked variation in both behavioral and neural responses independent of the presence of a signal (as evidenced by the absence of effects / interactions with stimulus type). This finding serves to motivate the application of reverse correlation to our data – if responses were mainly a function of signal presence (i.e., stimulus type) then CIs would reveal little if any information. And third, they suggest behaviorally and neurally-derived CIs may be similar to each other in virtue of the fact that they are constructed


John Wiley & Sons, Inc.
For Peer Review

from correlated signals: BOLD amplitudes tend to be higher on trials on which subjects classify the stimuli as 'faces'.

Finally, a critical assumption of reverse correlation methods is the consistency of the mechanism responsible for producing responses (e.g., the use of a same internal template) across extensive series of test sessions. Particularly problematic is the possibility of additional learning, change of strategy or, in our case, changes in neural perceptual processing. To verify this assumption we examined both behavioral markers (i.e., accuracy and bias) and BOLD responses (i.e., ROI-specific face selectivity) for the presence of consistent changes across sessions. Specifically, we computed Pearson correlations between each of these measures averaged within sessions and the corresponding session number. Our results showed no significant change in accuracy or bias for either participant ($p>0.10$). Similarly, no change in estimates of face selectivity was found for any of the ROIs examined ($p>0.10$).

*Behaviorally-derived CIs*

In order to assess the overall reliability and quality of the results, we examined the intermediate steps involved in the generation of the CIs. In theory, the most informative trials in a standard image classification paradigm should be those on which a subject responds incorrectly in that they reflect stronger reliance upon internal templates [Murray et al., 2002; Solomon, 2002]. Thus, we may expect that the contrast of the components based upon incorrect responses ($\mu_{NF}$ and $\mu_{FN}$) to be higher than that those based upon correct ones ($\mu_{FF}$ and $\mu_{NN}$) and, thus, to contribute more information to the construction of a CI. This expectation was borne out by the results of both subjects (Fig. 2). Specifically, the RMS contrast of the four image components showed higher levels for the two types of incorrect responses than for correct ones. In addition,


John Wiley & Sons, Inc.
For Peer Review

the contrast level of the resulting CIs was markedly larger than that of any single component suggesting that information combines (and is used) in a relatively consistent manner across the four types of trials. Thus, both subjects appear to make reliable and consistent use of internal templates.

While the results above point to the likely use of internal face templates, they do not speak to the structure or the nature of these templates. To deal with these issues, we conducted two sets of analyses, in the frequency domain and in the spatial domain, as detailed below.

First, in the frequency domain, we computed the squared amplitude energy of raw (i.e., unsmoothed) CIs as well as of the actual experimental stimuli (Fig. 6). Specifically, for each given image we computed the energy of a range of frequencies (in cycles / image) and averaged the results across orientations. In the case of experimental stimuli, this analysis examined whether their structure exhibited significant energy across the entire frequency band (as shown in Fig. 6a) and, thus, it provide subjects with the opportunity to exploit information from the range of frequencies most relevant for face detection (i.e., under 16 cycles / image) [Costen et al., 1996; Näsänen, 1999; Peli et al., 1994]. In the case of the CIs, the analysis examined the relative importance and use of different frequencies in our face detection task (Fig. 6b). In order to evaluate these latter results more rigorously, the analysis was repeated for an additional 100 CIs obtained by randomly permuting the behavioral responses of each participant. Overall, the comparison between actual and randomly derived CIs revealed higher amplitudes for the former. More importantly, behaviorally-derived CIs showed a marked decrease in amplitude across higher frequencies characteristic of natural images, and faces in particular [Keil, 2008]. In contrast, permutation-based CIs exhibited a roughly flat profile characteristic of the spectrum of white noise.

For Peer Review



John Wiley & Sons, Inc.

Second, in the spatial domain, smoothed CIs (Fig. 7a) were analyzed using a pixel test [Chauvin et al., 2005] with the goal of identifying areas of the image (i.e., pixels) whose luminance values differ significantly from chance. Of note, this image-based analysis focuses on low-frequency information (i.e., less than 8 cycles / image) both because it is better suited to deal with such information and because low-frequency energy dominates the spectral profile of the CIs (Fig. 6). The outcome of this analysis revealed a triangular pattern of dark regions corresponding roughly to the position of the eyes and the mouth. The images also displayed a markedly bright region corresponding to the upper brow.

Thus, the two sets of analyses concur on the presence of consistent visual structures used in face detection. More importantly, they identify the main spatial components of these structures associated with low-frequency information.

*Neurally-derived CIs*

Right-hemisphere ROIs showed higher face selectivity than their left homologues [Kanwisher et al., 1997; Puce et al., 1996] and, more critically, their selectivity was reliably replicated across sessions in both subjects (Table 1). Consequently, our analysis focuses on these ROIs; however, for completeness, CIs were separately constructed for all regions. Neurally-derived CIs were analyzed in the frequency domain and in the spatial domain following the same approach described above for behavioral CIs.

Out of all regions, the right FFA exhibited high amplitudes for a broad range of frequencies relative to baseline as well as an overall decrease in amplitude at higher frequencies.

For Peer Review



John Wiley & Sons, Inc.

Figure 6b-d displays these results for the right FFA and OFA and Supporting Information Figure 1 shows the results for their left homologues.

Some differences between the two participants are immediately apparent. For instance, we note that EC exhibits a better separation from baseline than EA in the case of the FFA. These differences are consistent both with behavioral performance (e.g., EA's bias for 'no face' responses) and neural response profiles (e.g., EA's smaller FFA face-selectivity – Table 1). In line with these differences, we expect EA's CIs to possess a smaller SNR than EC's. In the spatial domain, pixel tests confirmed this expectation in that EC's images displayed more extensive structures that EA's – see Fig. 7 and Supporting Information Fig. 2.

To boost the SNR of the present images and facilitate their interpretation we appealed to one simplifying assumption: facial symmetry. Given the sensitivity of face-selective regions to symmetry [Caldara and Seghier, 2009], it is plausible that some symmetrical features may be present in the internal template used for face detection. To examine this possibility, we averaged each CI with its mirror-symmetric version and submitted the results to a new set of analyses. This manipulation effectively doubles the number of trials used in constructing the raw images (since noise fields were independently generated for the right and the left sides) and is thus expected to increase their SNR [Murray et al., 2002]. Figure 8a displays the results for both behaviorally-derived CIs and right FFA CIs – unlike these CIs, those corresponding to the right OFA did not show any clear improvement over the initial results of the two subjects. The examination of the results shows that both eye-level regions and the mouth appear to serve as key elements for face detection.

Another facet of our investigation concerns the relationship between behaviorally and neurally-derived CIs. The standard analysis of BOLD amplitudes suggests that the two types of

For Peer Review



John Wiley & Sons, Inc.

CIs should be positively correlated with each other – insofar as 'face' responses are overall associated with higher ROI amplitudes (Fig. 5), this relationship presumably carries over to the CIs based upon behavioral and neural data. To evaluate this hypothesis we correlated smoothed symmetrical CIs (the left-hand half of each image). As expected, all pairs of CIs showed positive correlations with each other: behavioral and FFA-based CIs (EC: $r=0.35$, $p<0.001$; EA: $r=0.26$, $p<0.001$) as well as behavioral and OFA-based CIs (EC: $r=0.28$, $p<0.001$; EA: $r=0.27$, $p<0.001$) – the overlap of significant regions contained by the two types of CIs is also shown in Supporting Information Figure 3. Similar correlation results were obtained by comparing smoothed CIs prior to introducing a symmetry assumption. However, this time the correlation between the right FFA and the behavioral CIs derived for subject EA did not reach significance ($p > 0.10$) – this latter result is consistent with the standard analyses mentioned above given that, unlike the other ROIs examined, EA's FFA did not show a significant effect of response type (i.e., 'face' versus 'noise').

Correlations between behavioral and BOLD of data are certainly important in clarifying the relationship between brain and behavior. At the same time though, if neurally-derived CIs display significant structures simply by virtue of the correlation with behavior, in the long term, the application of image classification to BOLD data may prove to be of limited theoretical value. To address this concern we computed and analyzed a new set of neurally-based CIs. Specifically, first, we regressed out behavioral responses from neural ones and, second, we constructed CIs from the neural residuals of both the right FFA and the right OFA of each subject. Image-based analyses show that the new CIs (Supporting Information Fig. 4) exhibit most of the significant regions present in the original CIs (Fig. 7b-d). Therefore, we argue, the correlation of BOLD and behaviorally responses is not the only (or even the main) factor


John Wiley & Sons, Inc.
For Peer Review

responsible for the structure of neurally-based CIs. More generally, this latter result supports the idea that BOLD-derived CIs can contribute significant information regarding visual representations independent of that provided by their behavioral counterparts.

## Discussion

*Internal face representations*

What basic image structures guide face detection within the human visual system? Our study uses image classification to clarify the structure of general face representations and their instantiation at the neural level (i.e., at the level of face-selective regions). Overall, our results provide evidence for simple but robust image structures including a triangular configuration of dark areas corresponding to the eyes and the mouth along with brighter areas corresponding to the middle brow. These image structures are especially clear based on our behavioral results; however, their elements can be traced to neural processing, especially in the case of the right FFA.

The present results are in broad agreement with recent behavioral and EEG studies [Hansen et al., 2010; Rieth et al., 2011; Smith et al., 2012] that also reveal significant image structures mediating face detection**. For instance, Rieth and colleagues [2011] applied image classification to behavioral data collected across a large number of subjects (i.e., several hundred). The resultant CIs associated with a face detection task showed a multitude of dark patches across a surprisingly wide expanse of the image both centrally and peripherally – however, guiding the subjects' attention to the center of the image reduced the amount of spatial uncertainty leading to a less dispersed and more intuitive 'face-like'**



John Wiley & Sons, Inc.

For Peer Review

**structure. More relevantly here, two other studies [Hansen et al., 2010; Smith et al., 2012] extended image classification to neuroimaging data associated with a face detection task. Specifically, these studies derived CIs corresponding to EEG signals at different time points and frequency bands. In one study, significant image structures were observed in multiple frequency bands for occipitotemporal cortex around 170 ms [Hansen et al., 2010]. While these structures were fairly diverse in their appearance across bands and subjects, additional analyses across CIs confirmed that they were likely to contain visual features characteristic of actual faces. Interestingly, Smith et al. [2012] showed that meaningful structures can be derived from frontal areas as well as occipitotemporal areas in a broad interval ranging from 200 to 500ms from stimulus onset. Moreover, neurally-derived CIs correlated reliably in this latter study with their behavioral counterparts computed across the same subjects, thus reinforcing the explanatory value of the neural results.**

Importantly, some of the image structures identified by the studies above have noticeable similarity to those we found here – for instance, the eyes appear to play a dominant role. Thus, the successful application of image classification across neuroimaging modalities (i.e., EEG, by previous studies, and fMRI here) suggests that our current results reflect meaningful aspects of neural representations. At the same time, as expected, we note the presence of substantial variability in the overall pattern of results across subjects (and across studies). Thus, our results also underline the clear challenges facing further applications of image classification to neural data (see next section). Moreover, **while the studies above argue for the intuitive 'face-like' aspects of certain image structures**, it remains unclear what properties recommend these structures for their privileged role in recognition and for their encoding in high-level visual areas. That is, the identification of significant image structures can benefit from an explanation of their


John Wiley & Sons, Inc.
For Peer Review

function. In this respect, we argue that a plausible account of the present results involves objective feature diagnosticity as detailed below.

Previous investigations of visual face properties found that the contrast between the eyes and the rest of the face is highly characteristic of faces as a visual category [Gilad et al., 2009; Sinha, 2002]. In particular, the contrast between the eyes and the middle brow area or the upper cheeks (Fig. 8b) systematically outperforms other local features in automatic recognition [Viola and Jones, 2004]. Consistent with this, the horizontal placement of the internal features (e.g., the eyes and the middle brow) leads to a specific face signature in the frequency domain – the presence of diagnostic information at around 10 cycles per face [Keil, 2008]. Furthermore, recent comparisons of automatic and human face detection ['t Hart et al., 2011] suggest that simple contrast features such as those in Figure 8b are highly predictive of behavioral performance. To be clear, such features are not invariant (e.g., an extreme change in viewpoint can render them relatively ineffective). However, what matters is their robustness over a large number of common changes, both intrinsic (e.g., expression) and extrinsic (e.g., lighting). Overall, our results provide support for these previous findings by deriving such features directly from patterns of behavioral and neural responses. Conversely, these previous findings support the idea that the most robustly encoded features are those most diagnostic about faces as a class. In this sense, face encoding appears to reflect the objective structure and statistics of face images [Bartlett, 2007] in a manner that is similar to the way early visual representations reflect the low-level statistics of natural images [Barlow, 1961; Olshausen and Field, 1996].

As far as the frequency profile of the features noted above is concerned, their coarse low-resolution aspect (under 8 cycles / image) is quite obvious. This result may seem at odds with the availability of high-frequency information for detection purposes [Halit et al., 2006]. However,

For Peer Review



John Wiley & Sons, Inc.

current research suggests that high-frequency information is not critical for face recognition [Costen et al., 1996; Näsänen, 1999; Peli et al., 1994]. Face detection as carried out by the human visual system is remarkably fast and efficient, for instance when compared with individuation [Liu et al., 2002; Or and Wilson, 2010]. As such, it is likely to take advantage more readily of low-frequency information whose availability precedes that of high-frequency information [Bar et al., 2006]. Thus, the privileged role of low frequencies in neural processing along with the diagnosticity of the information they carry serve as a plausible explanation for the coding of the features revealed by our CIs.

Interestingly, the structures revealed by our CIs bears similarity to the type of simple displays (e.g., low-frequency eyes-and-mouth configurations) evoking preferential looking in infants [Farroni et al., 2005; Johnson and Morton, 1991]. This structure has been associated in the past with subcortical face processing [Johnson, 2005]. While our results do not speak directly to this possibility, the diagnosticity of the visual features identified provides a plausible argument for their redundant encoding at multiple levels of visual processing. In particular, we find that the right FFA appears to encode these features confirming its involvement in face detection [Freiwald et al., 2009; Grill-Spector et al., 2004; Nestor et al., 2008].

The considerations above raise an interesting issue: to what extent the study of other types of face stimuli (e.g., profiles) or even other categories of objects would reveal significant structures such as those found here in the FFA? To address this issue three related points should to be considered. First, face profiles activate the FFA significantly less than frontal-view faces [Xu et al., 2009; Yue et al., 2011] and so do objects [Kanwisher and Yovel, 2006]. Second, face profile and object features encoded in the FFA are probably less diagnostic for their respective classes and, therefore, less robustly encoded – for instance, highly effective features for face


John Wiley & Sons, Inc.
For Peer Review

detection like those shown in Fig. 8b have difficulty in dealing with profiles. Third, the FFA does not appear to respond linearly to other categories than faces [Horner and Andrews, 2009] warning against the application of standard image classification to such cases. Therefore, we argue, the investigation of alternative types of stimuli in face-selective regions is likely to be less informative. At the same time though, we do acknowledge that such investigations may serve as relevant controls in the evaluation of image classification results such as those presented here.

On a related note, it may seem surprising to assume linearity in the neural responses of high-level visual areas associated with any object category, particularly considering that invariance in object recognition is achieved primarily trough nonlinear processing [Riesenhuber and Poggio, 1999]. **However, such nonlinearities in face processing may be amenable to linear approximations under certain respects.** Indeed, recent evidence suggests that the FFA exhibits much less invariance to basic image characteristics than previously thought. For instance, its response increases with the size of a face stimulus [Xu et al., 2009; Yue et al., 2011], decreases with its eccentricity [Schwarzlose et al., 2008; Yue et al., 2011] and also with viewpoint divergence from a frontal view [Xu et al., 2009; Yue et al., 2011]. Furthermore, many of these properties affect response amplitudes and combine with each other in a roughly linear fashion [Yue et al., 2011]. Finally, the response of the FFA was found to increase proportionately with the 'faceness' of a stimulus [Davidenko et al., 2011; Horner and Andrews, 2009]. **To be clear, these results do not imply that the functioning of the FFA reduces to strictly linear operations but rather that important aspects of its functioning, such as those related to face detection, can be reasonably approximated by a linear function**. Thus, the response characteristics of the FFA make it ideal for the goals of our investigation while they

For Peer Review



John Wiley & Sons, Inc.

also raise questions concerning the more general applicability of image classification to BOLD data as discussed in the next section.

In sum, our results argue for the role of several critical features in face detection. Clearly though, face detection is not limited to their use. For instance, other luminance-based features (e.g., hair), although less stable and robust, are likely to complement those discussed here. Similarly, other modalities in addition to luminance can provide diagnostic information. As a case in point, color can be exploited in a number of face recognition tasks including detection [Bindemann and Burton, 2009; Dupuis-Roy et al., 2009; Nestor and Tarr, 2008]. Thus, we argue that the features discussed here serve as robust properties of face representations rather than as complete and flawless ones.

Finally, as a point of clarification, we note that top-down processes such as expectation and context are unlikely to account for the present results. The use of noisy / ambiguous images is a powerful tool for researching top-down processes in object recognition [Li et al., 2009; Summerfield et al., 2006; Wild and Busey, 2004]. The general strategy of this research involves direct pairing of neural responses with higher-level cognitive factors (e.g., expectations regarding probability of occurrence). In contrast, image classification as illustrated here aims at relating behavioral / cortical responses with random image structures. The relevant factor in this relationship is the accidental similarity of these structures to actual face representations rather than any manipulation of high-level cognitive factors. Thus, we argue that our results serve as approximations of internal visual representations rather than as byproducts of top-down visual processing.

For Peer Review



John Wiley & Sons, Inc.

*Application of image classification to fMRI*

The present findings suggest that an extension of noise-based image classification to BOLD data can be informative as long as several preconditions are satisfied. First, the overall linearity of response amplitudes within a region [Davidenko et al., 2011; Horner and Andrews, 2009; Yue et al., 2011] is likely to be an important factor in this respect. Second, the systematic variability of neural responses (e.g., as illustrated by their relationship with behavioral responses) is critical to constructing meaningful CIs. Third, optimizing the SNR of the CIs serves as a significant constraint both in the design of the experimental paradigm and in the construction of the CIs. Given such considerations, our results provide a proof of principle that image classification can be applied to BOLD data to uncover visual features employed in high-level recognition.

At the same time, we note that the completeness and quality of neurally-based CIs is not on the same par with that of behaviorally-derived CIs as illustrated by the present results and by related studies [Hansen et al., 2010; Smith et al., 2012]. At least two reasons seem to underlie this difference. First, the SNR of neural recordings is likely poorer than that of behavioral responses. For instance, BOLD signals are corrupted both by internal (e.g., physiological) noise and by external noise related to fMRI measurement (e.g., thermal noise) [Bennett and Miller, 2010]. In contrast, behavioral responses are primarily influenced by internal noise – there is virtually no measurement noise associated with recording button presses. Our work attempts to deal with this issue by maximizing the SNR of BOLD-derived CIs. Despite such efforts, it seems unlikely that current methods can yield comparable SNR levels for the two categories of data. Second, it is reasonable to assume that any CI based upon activation in a single brain region may

For Peer Review



John Wiley & Sons, Inc.

provide only a noisy and incomplete estimation of the overall internal template driving behavioral responses. For instance, in the case of face perception, its reliance upon an entire network of cortical regions is well-documented [Gauthier et al., 2000; Haxby et al., 2000; Haxby et al., 2001; Ishai et al., 2005; Rossion et al., 2003; Tsao et al., 2008] and consistent with the idea that these regions provide both redundant and complementary information for the purpose of face recognition [Fox et al., 2009; Gobbini and Haxby, 2007; Nestor et al., 2011]. Thus, the construction of hybrid CIs based from patterns of activation across multiple regions may ultimately provide a way to boost the quality of neurally-derived CIs. More generally, relating and combining CIs across different brain regions in a principled and statistically optimal manner may provide new insights into how information is integrated at the level of cortical networks and how behavior emerges from complex visual processing.

In addition to the research directions noted above, a more extensive application of image classification to neuroimaging, and BOLD data in particular, appears to require two critical developments. One concerns a significant reduction in the number of trials, for instance by replacing random sampling with adaptive stimulus sampling [Lewi et al., 2009]. Generating and testing maximally informative stimuli on the fly as a function of previous responses is certainly an option for behavioral studies but also for neuroimaging, particularly in connection with the advent of real-time fMRI [deCharms, 2008; LaConte et al., 2007]**.** The other development involves the use of nonlinear methods [Neri, 2004] better suited to uncovering subtler, more complex, higher-level features. Such developments are critical in extending the application of image classification beyond the interesting but restricted domain of face detection.

At this time, we note that neuroimaging data represent a new domain for the application of image classification. 'Bubbles' [Gosselin and Schyns, 2001], a technique related to image


John Wiley & Sons, Inc.
For Peer Review

classification, has been recently applied to several imaging modalities including fMRI [Smith et al., 2007; Smith et al., 2008; Smith et al., 2009]. However, rather than aiming to reconstruct internal representations out of structure-free stimuli, the bubbles technique takes on the interesting but less taxing enterprise of uncovering informative areas of a given stimulus. More closely related to our work, another study [Smith et al., 2012] applied a challenging version of image classification known as 'superstitious perception' [Gosselin and Schyns, 2003] to EEG data. Unlike standard image classification [Abbey and Eckstein, 2002; Ahumada, 2002; Beard and Ahumada, 1998], superstitious perception completely discards the use of a base image and only relies on noise stimuli to construct CIs. The merit of this approach is obvious in that it forces a heavier reliance on internal templates in performing the task. However, it also introduces the risk of variable / evolving internal templates within (and across) subjects, a risk that base images, such as those used here, are intended to minimize. From a practical point of view, this version of image classification may not be immediately applicable to BOLD data due to the larger number of trials needed. However, the development of adaptive stimulus sampling [Lewi et al., 2009] in connection with real-time fMRI could make superstitious perception a feasible and appealing approach for future research.

Finally, an interesting parallel can be drawn here with fMRI methods for stimulus reconstruction [Miyawaki et al., 2008; Naselaris et al., 2009]. The idea of reconstructing an image-based structure is common to both such methods and to reverse correlation. However, the general goal of the former is to reconstruct actual stimuli from neural patterns rather than to recover the structure of neural representation. Thus, while impressive as an engineering feat, stimulus reconstruction was pointed out to have unclear theoretical value [Kriegeskorte, 2011] in that it exploits current knowledge about neural representations rather than attempting to extend

For Peer Review



John Wiley & Sons, Inc.

it. In particular, stimulus reconstruction takes advantage of existing computational descriptions of neural representations in early visual areas. On the other hand, rigorous descriptions at the level of higher visual areas are still missing. In this respect, image classification methods may provide an important tool by narrowing the gap between models of neural representation at the level of high-level versus low-level visual areas.

*Summary*

Our work aims at uncovering the basic visual structures underlying human face detection and at relating them to the neural representations hosted by ventral face-selective areas. Our results reveal the existence and characteristics of such structures and account for them in terms of their objective diagnosticity for face detection. More generally, the present results are instrumental in establishing the potential as well as the challenges confronting the application of image classification to BOLD data in the study of high-level visual perception.

For Peer Review



John Wiley & Sons, Inc.

**Acknowledgements**. We thank David Badre, David Sheinberg and Isabel Gauthier for valuable and helpful comments on previous versions of this paper. This work was funded by NIH EUREKA award #1R01MH084195-01, a James S. McDonnell Foundation grant to the Perceptual Expertise Network (PEN) and an NSF Science of Learning Center grant SBE-0542013 to the Temporal Dynamics of Learning Center (TDLC).

For Peer Review



John Wiley & Sons, Inc.

## References

Abbey CK, Eckstein MP (2002) Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. J Vis 2:66-78.

Ahumada AJ (2002) Classification image weights and internal noise level estimation. J Vis 2:121-131.

Avidan G, Hasson U, Malach R, Behrmann M (2005) Detailed exploration of face-related processing in congenital prosopagnosia: 2. Functional neuroimaging findings. J Cogn Neurosci 17:1150-1167.

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Sensory Communication (Rosenblith W, eds), pp 217-234. Cambridge, MA: MIT Press.

Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Schmidt AM, Dale AM, Hämäläinen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E (2006) Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A 103:449-454.

Bartlett MS (2007) Information maximization in face processing. Neurocomputing 70:2204-2217.

Beard BL, Ahumada AJ (1998) A technique to extract the relevant features for visual task. In: Human vision and electronic imaging III (SPIE Proceedings Vol. 3299) (Rogowitz BE, Pappas TN, eds), pp 79-85. Bellingham, WA: International Society for Optical Engineering.

Bennett CM, Miller MB (2010) How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 1191:133-155.

For Peer Review



John Wiley & Sons, Inc.

Bindemann M, Burton AM (2009) The role of color in human face detection. Cogn Sci 33:1144-1156.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433-436.

Caldara R, Seghier ML (2009) The Fusiform Face Area responds automatically to statistical regularities optimal for face categorization. Hum Brain Mapp 30:1615-1625.

Chauvin A, Worsley KJ, Schyns PG, Arguin M, Gosselin F (2005) Accurate statistical tests for smooth classification images. J Vis 5:659-667.

Costen NP, Parker DM, Craw I (1996) Effects of high-pass and low-pass spatial filtering on face identification. Percept Psychophys 58 :602-612.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162-173.

Davidenko N, Remus DA, Grill-Spector K (forthcoming) Face-likeness and image variability drive responses in human face-selective ventral regions. Hum Brain Mapp.

deCharms RC (2008) Applications of real-time fMRI. Nat Rev Neurosci 9:720-729.

Dupuis-Roy N, Fortin I, Fiset D, Gosselin F (2009) Uncovering gender discrimination cues in a realistic setting. J Vis 9:10.1-10.8.

Farroni T, Johnson MH, Menon E, Zulian L, Faraguna D, Csibra G (2005) Newborns' preference for face-relevant stimuli: effects of contrast polarity. Proc Natl Acad Sci U S A 102:17245-17250.

Fox CJ, Moon SY, Iaria G, Barton JJ (2009) The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. Neuroimage 44:569-580.

Freiwald WA, Tsao DY, Livingstone MS (2009) A face feature space in the macaque temporal lobe. Nat Neurosci 12:1187-1196.

For Peer Review



John Wiley & Sons, Inc.

Friston KJ, Jezzard P, Turner R (1994) Analysis of functional MRI time-series. Hum Brain Mapp 1:153-171.

Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JC, Anderson AW (2000) The fusiform "face area" is part of a network that processes faces at the individual level. J Cogn Neurosci 12:495-504.

Gilad S, Meng M, Sinha P (2009) Role of ordinal contrast relationships in face encoding. Proc Natl Acad Sci U S A 106:5353-5358.

Gobbini MI, Haxby JV (2007) Neural systems for recognition of familiar faces. Neuropsychologia 45:32-41.

Gold JM, Murray RF, Bennett PJ, Sekuler AB (2000) Deriving behavioural receptive fields for visually completed contours. Curr Biol 10:663-666.

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res 41:2261-2271.

Gosselin F, Schyns PG (2003) Superstitious perceptions reveal properties of internal representations. Psychol Sci 14:505-509.

Grill-Spector K, Knouf N, Kanwisher N (2004) The fusiform face area subserves face perception, not generic within-category identification. Nat Neurosci 7:555-562.

Halit H, de Haan M, Schyns PG, Johnson MH (2006) Is high-spatial frequency information used in the early stages of face detection? Brain Res 1117:154-161.

Hansen BC, Thompson B, Hess RF, Ellemberg D (2010) Extracting the internal representation of faces from human brain activity: an analogue to reverse correlation. Neuroimage 51:373-390.

For Peer Review



John Wiley & Sons, Inc.

Harel A, Ullman S, Epshtein B, Bentin S (2007) Mutual information of image fragments predicts categorization in humans: Electrophysiological and behavioral evidence. Vision Res 47:2010-2020.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223-233.

Horner AJ, Andrews TJ (2009) Linearity of the fMRI response in category-selective regions of human visual cortex. Hum Brain Mapp.

Ishai A, Schmidt CF, Boesiger P (2005) Face perception is mediated by a distributed cortical network. Brain Res Bull 67:87-93.

Jiang Y, He S (2006) Cortical responses to invisible faces: dissociating subsystems for facial-information processing. Curr Biol 16:2023-2029.

Johnson MH (2005) Subcortical face processing. Nat Rev Neurosci 6:766-774.

Johnson MH, Morton J (1991) Biology and Cognitive Development: The Case of Face Recognition. Oxford: Blackwell.

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

Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci 361:2109-2128.

Keil MS (2008) Does face image statistics predict a preferred spatial frequency for human face processing? Proc Biol Sci 275:2095-2100.

For Peer Review



John Wiley & Sons, Inc.

Kriegeskorte N (2011) Pattern-information analysis: from stimulus decoding to computational-model testing. Neuroimage 56:411-421.

Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A 104:20600-20605.

Kriegeskorte N, Lindquist MA, Nichols TE, Poldrack RA, Vul E (2010) Everything you never wanted to know about circular analysis, but were afraid to ask. J Cereb Blood Flow Metab 30:1551-1557.

LaConte SM, Peltier SJ, Hu XP (2007) Real-time fMRI using brain-state classification. Hum Brain Mapp 28:1033-1044.

Lewi J, Butera R, Paninski L (2009) Sequential optimal design of neurophysiology experiments. Neural Comput 21:619-687.

Li J, Liu J, Liang J, Zhang H, Zhao J, Huber DE, Rieth CA, Lee K, Tian J, Shi G (2009) A distributed neural system for top-down face processing. Neurosci Lett 451:6-10.

Liu J, Harris A, Kanwisher N (2002) Stages of processing in face perception: an MEG study. Nat Neurosci 5:910-916.

Loffler G, Yourganov G, Wilkinson F, Wilson HR (2005) fMRI evidence for the neural representation of faces. Nat Neurosci 8:1386-1390.

Miyawaki Y, Uchida H, Yamashita O, Sato MA, Morito Y, Tanabe HC, Sadato N, Kamitani Y (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60:915-929.

Murray RF, Bennett PJ, Sekuler AB (2002) Optimal methods for calculating classification images: Weighted sums. J Vis 2:79-104.

For Peer Review



John Wiley & Sons, Inc.

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63:902-915.

Näsänen R (1999) Spatial frequency bandwidth used in the recognition of facial images. Vision Res 39:3824-3833.

Neri P (2004) Estimation of nonlinear psychophysical kernels. J Vis 4:82-91.

Neri P, Levi DM (2006) Receptive versus perceptive fields from the reverse-correlation viewpoint. Vision Res 46:2465-2474.

Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc Natl Acad Sci U S A 108:9998-10003.

Nestor A, Tarr MJ (2008) Gender recognition of human faces using color. Psychol Sci 19:1242-1246.

Nestor A, Vettel JM, Tarr MJ (2008) Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. PLoS ONE 3:e3978.

Olshausen BA, Field DJ (1996) Natural image statistics and efficient coding. Network 7:333-339.

Or CC, Wilson HR (2010) Face recognition: Are viewpoint and identity processed after face detection? Vision Res 50:1581-1589.

Peli E, Lee E, Trempe CL, Buzney S (1994) Image enhancement for the visually impaired: the effects of enhancement on face recognition. J Opt Soc Am A Opt Image Sci Vis 11:1929-1939.

Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat Vis 10:437-442.

For Peer Review



John Wiley & Sons, Inc.

Pourtois G, Schwartz S, Seghier ML, Lazeyras F, Vuilleumier P (2005) View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: an event-related fMRI study. Neuroimage 24:1214-1224.

Puce A, Allison T, Asgari M, Gore JC, McCarthy G (1996) Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. J Neurosci 16:5205-5215.

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019-1025.

Rieth CA, Lee K, Lui J, Tian J, Huber DE (2011) Faces in the mist: Illusory face and letter detection. i-Perception 2:458-476.

Righart R, Andersson F, Schwartz S, Mayer E, Vuilleumier P (2010) Top-down activation of fusiform cortex without seeing faces in prosopagnosia. Cereb Cortex 20:1878-1890.

Ringach D, Shapley R (2004) Reverse correlation in neurophysiology. Cogn Sci 28:147-166.

Rossion B, Caldara R, Seghier M, Schuller AM, Lazeyras F, Mayer E (2003) A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. Brain 126:2381-2395.

Schwarzlose RF, Swisher JD, Dang S, Kanwisher N (2008) The distribution of category and location information across object-selective regions in human visual cortex. Proc Natl Acad Sci U S A 105:4447-4452.

Sinha P (2002) Qualitative representations for recognition. In: Lecture Notes in Computer Science (Springer-Verlag, eds), pp 249-262. .

For Peer Review



John Wiley & Sons, Inc.

Smith FW, Muckli L, Brennan D, Pernet C, Smith ML, Belin P, Gosselin F, Hadley DM, Cavanagh J, Schyns PG (2008) Classification images reveal the information sensitivity of brain voxels in fMRI. Neuroimage 40:1643-1654.

Smith ML, Fries P, Gosselin F, Goebel R, Schyns PG (2009) Inverse mapping the neuronal substrates of face categorizations. Cereb Cortex 19:2428-2438.

Smith ML, Gosselin F, Schyns PG (2007) From a face to its category via a few information processing states in the brain. Neuroimage 37:974-984.

Smith ML, Gosselin F, Schyns PG (2012) Measuring internal representations from behavioral and brain data. Curr Biol 22:191-196.

Solomon JA (2002) Noise reveals visual mechanisms of detection and discrimination. J Vis 2:105-120.

Summerfield C, Egner T, Mangels J, Hirsch J (2006) Mistaking a house for a face: neural correlates of misperception in healthy humans. Cereb Cortex 16:500-508.

't Hart BM, Abresch TG, Einhäuser W (2011) Faces in places: humans and machines make similar face detection errors. PLoS ONE 6:e25373.

Tong F, Nakayama K, Moscovitch M, Weinrib O, Kanwisher N (2000) Response properties of the human fusiform face area. Cogn Neuropsychol 17:257-280.

Tsao DY, Livingstone MS (2008) Mechanisms of face perception. Annu Rev Neurosci 31:411-437.

Tsao DY, Moeller S, Freiwald WA (2008) Comparing face patch systems in macaques and humans. Proc Natl Acad Sci U S A, 105:19513-19-518

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5:682-687.

For Peer Review



John Wiley & Sons, Inc.

Viola P, Jones MJ (2004) Robust real-time face detection. International Journal of Computer Vision 57:137-154.

Watson AB, Pelli DG (1983) QUEST- A Bayesian adaptive psychometric method. Perception and Psychophysics 33:113-120.

Wild HA, Busey TA (2004) Seeing faces in the noise: stochastic activity in perceptual regions of the brain may influence the perception of ambiguous stimuli. Psychon Bull Rev 11:475-481.

Winston JS, Henson RN, Fine-Goulden MR, Dolan RJ (2004) fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J Neurophysiol 92:1830-1839.

Xu X, Yue X, Lescroart MD, Biederman I, Kim JG (2009) Adaptation in the fusiform face area (FFA): image or person? Vision Res 49:2800-2807.

Yue X, Cassidy BS, Devaney KJ, Holt DJ, Tootell RB (2011) Lower-level stimulus features strongly influence responses in the fusiform face area. Cereb Cortex 21:35-47.

Zhang H, Liu J, Huber DE, Rieth CA, Tian J, Lee K (2008) Detecting faces in pure noise images: a functional MRI study on top-down perception. Neuroreport 19:229-233.

For Peer Review



John Wiley & Sons, Inc.

**Figure Captions**

Fig 1. Base image (left) and examples of the two types of stimuli presented in the experiment: noise-only (middle) and noise-plus-base image (right)

Fig 2. Intermediate results involved in the construction of behaviorally-derived CIs: four different components corresponding to four types of trials are added to each other in order to estimate the internal template guiding behavioral responses (note that the polarity of $\mu_{FN}$ and $\mu_{NN}$ was flipped for ease of visualization and comparison with the other components). Smoothed components and CIs are displayed under their original (raw) versions. The RMS contrast of each raw image is separately computed for each component and CI. Results are separately shown for subjects EC (top) and EA (bottom).

Fig 3. Procedure for the construction of neurally-derived CIs: (a) noise fields are grouped and averaged based on ROI response amplitude and stimulus type (F – face base image, N – noise) at a given time point; (b) noise field averages are weighted and combined into a time-specific CI; (c) a weighted sum is computed across time-specific CIs using a standard hemodynamic response function (HRF) to generate a single ROI-specific CI; (d) raw CIs are smoothed with a Gaussian filter to allow analysis and visualization.

Fig 4. Example of ROI mask in subject EA. The map shows the contrast between faces and objects ($q$<0.05) superimposed on three axial slices (in EA's native space). The mask is centered on the peak of the right FFA (see Table 1).

For Peer Review



John Wiley & Sons, Inc.

Fig 5. Response amplitudes (in percent signal change) across different ROIs as a function of stimulus type and behavioral response (h – hits, fa – false alarms, m – misses, cr – correct rejections). Error bars show ±1SE across sessions.

Fig 6. Average squared amplitude energy for (a) the base image, stimuli containing the base image and stimuli containing only noise fields; (b) the raw behavioral CIs; (c), (d) the raw neurally-derived CIs (corresponding to the right FFA and OFA). The abscissa represents spatial frequency in cycles per image and the ordinate displays normalized amplitude values averaged across orientations – values are normalized (scaled) by the maximum value. The average energy of 100 control CIs (constructed by permuting response labels) is shown in gray. Error bars show ±1SD across stimuli for (a) and across control CIs for (b-d).

Fig 7. Smoothed CIs and their pixel test analysis. Results are shown for (a) behavioral responses, (b) right FFA responses and (c) the right OFA responses. Blue and yellow mark pixels darker / brighter than chance ($p$<0.05).

Fig 8. (a) Symmetrical CIs analyzed with a pixel test ($p$<0.05). Results are shown for behavioral CIs (on the left) and for rFFA-derived CIs (on the right). (b) The two best contrast features for face detection of Viola and Jones [2004] superimposed on a base image.

SI Fig 1. Average squared amplitude energy for raw neurally-derived CIs. Results are shown for the left FFA and OFA. The abscissa represents spatial frequency and the ordinate displays normalized amplitude values (averaged across orientations). The average energy of 100 control CIs (constructed by permuting ROI response amplitudes) is shown in gray (±1SD).

For Peer Review



John Wiley & Sons, Inc.

SI Fig 2. Smoothed neurally-derived CIs analyzed with a pixel test. Results are shown for the left FFA (on the left) and the left OFA (on the right). Blue and yellow mark pixels darker / brighter than chance ($p<0.05$).

SI Fig 3. Overlap of significant regions in behaviorally and neurally-derived symmetrical CIs for both subjects: behavioral – right FFA CI overlap (left column) and behavioral – right OFA CI overlap (right column). Blue and yellow mark pixels darker / brighter than chance (significance levels for each of the conjuncts were set to $p<0.10$ in order to allow a more thorough examination of the overlap). Overlapping regions are superimposed on a base image to aid visualizing the spatial correspondence to actual facial features.

SI Fig 4. Smoothed neurally-derived CIs and their pixel test analysis. CIs were computed based on the residuals of ROI-specific response amplitudes after factoring out behavioral responses. Results are shown for (a) the right FFA and (b) the right OFA. Blue and yellow mark pixels darker / brighter than chance ($p<0.05$).

**Table captions**

Table 1. Coordinates and average face selectivity for the ROIs (EC/EA). Face selectivity is measured as the difference between face and object-evoked activation (* $p<0.05$, ** $p<0.001$).

For Peer Review



John Wiley & Sons, Inc.

| ROI | Peak coordinates | | | Face selectivity (%SC ±SD) |
|---|---|---|---|---|
| | x | y | z | |
| rFFA | 45 / 45 | -53 / -44 | -15 / -18 | 0.26 (±0.15)** / 0.16 (±0.08)** |
| lFFA | -40 / -38 | -46 / -40 | -11 / -14 | 0.05 (±0.05)* / 0.05 (±0.08) |
| rOFA | 36 / 41 | -77 / -76 | -12 / -5 | 0.22 (±0.11)** / 0.30 (±0.13)** |
| lOFA | -39 / -41 | -78 / -73 | -11 / -5 | 0.09 (±0.19) / 0.22 (±0.12)** |
| EVC | 16 / 1 | -94 / -93 | -6 /-7 | 0.01 (±0.17) / 0.07 (±0.15) |

For Peer Review

Fig 1. Base image (left) and examples of the two types of stimuli presented in the experiment: noise-only (middle) and noise-plus-base image (right)
32x11mm (300 x 300 DPI)

For Peer Review

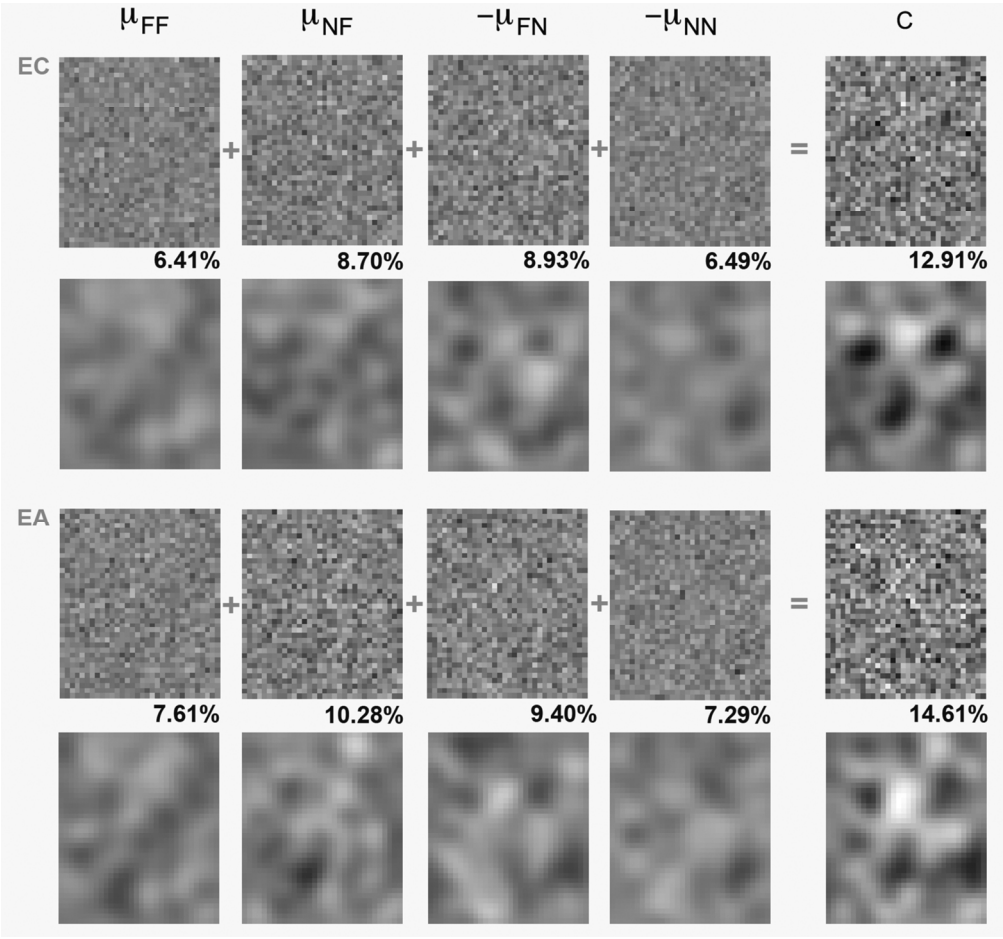


Fig 2. Intermediate results involved in the construction of behaviorally-derived CIs: four different components corresponding to four types of trials are added to each other in order to estimate the internal template guiding behavioral responses (note that the polarity of and was flipped for ease of visualization and comparison with the other components). Smoothed components and CIs are displayed under their original (raw) versions. The RMS contrast of each raw image is separately computed for each component and CI. Results are separately shown for subjects EC (top) and EA (bottom).
139x130mm (300 x 300 DPI)

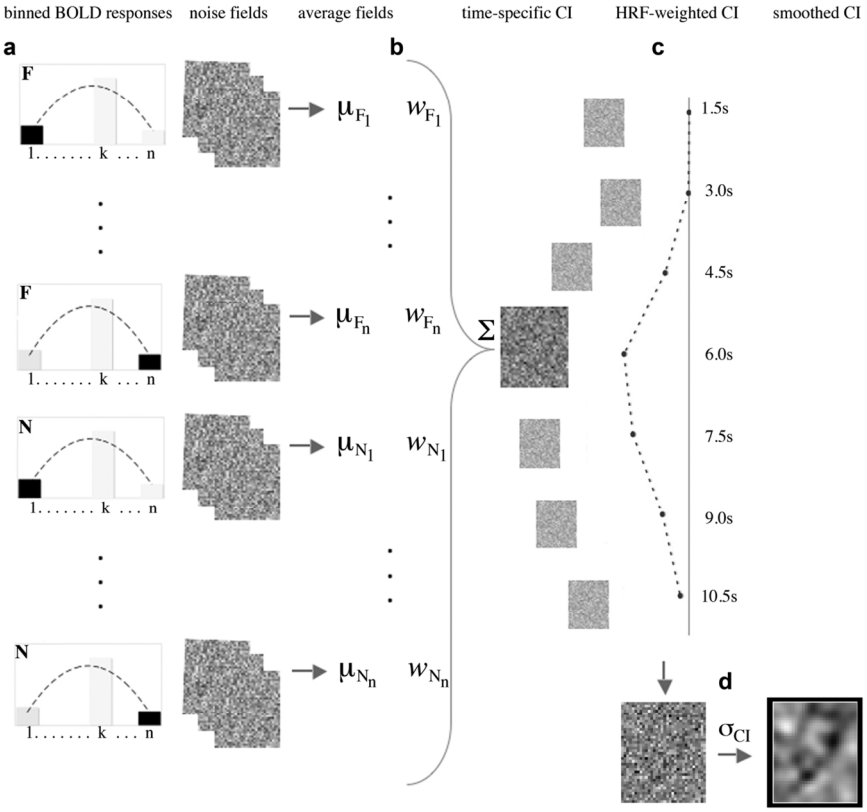


Fig 3. Procedure for the construction of neurally-derived CIs: (a) noise fields are grouped and averaged based on ROI response amplitude and stimulus type (F – face base image, N – noise) at a given time point; (b) noise field averages are weighted and combined into a time-specific CI; (c) a weighted sum is computed across time-specific CIs using a standard hemodynamic response function (HRF) to generate a single ROI-specific CI; (d) raw CIs are smoothed with a Gaussian filter to allow analysis and visualization.
134x110mm (300 x 300 DPI)

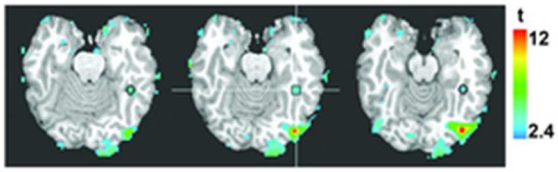


Fig 4. Example of ROI mask in subject EA. The map shows the contrast between faces and objects ($q<0.05$) superimposed on three axial slices (in EA's native space). The mask is centered on the peak of the right FFA (see Table 1).
25x7mm (300 x 300 DPI)

For Peer Review

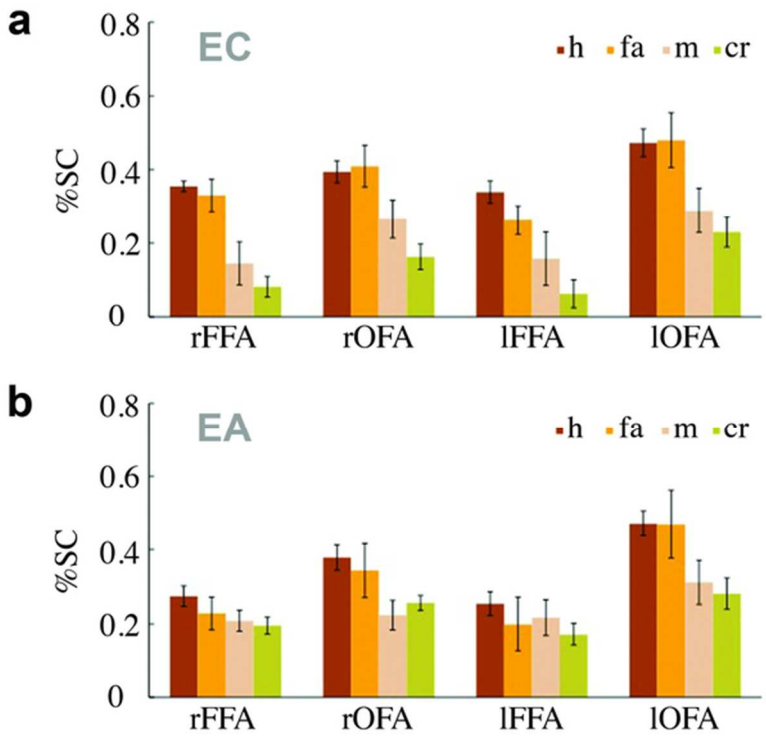


Fig 5. Response amplitudes (in percent signal change) across different ROIs as a function of stimulus type and behavioral response (h – hits, fa – false alarms, m – misses, cr – correct rejections). Error bars show ±1SE across sessions.
66x49mm (300 x 300 DPI)

For Review

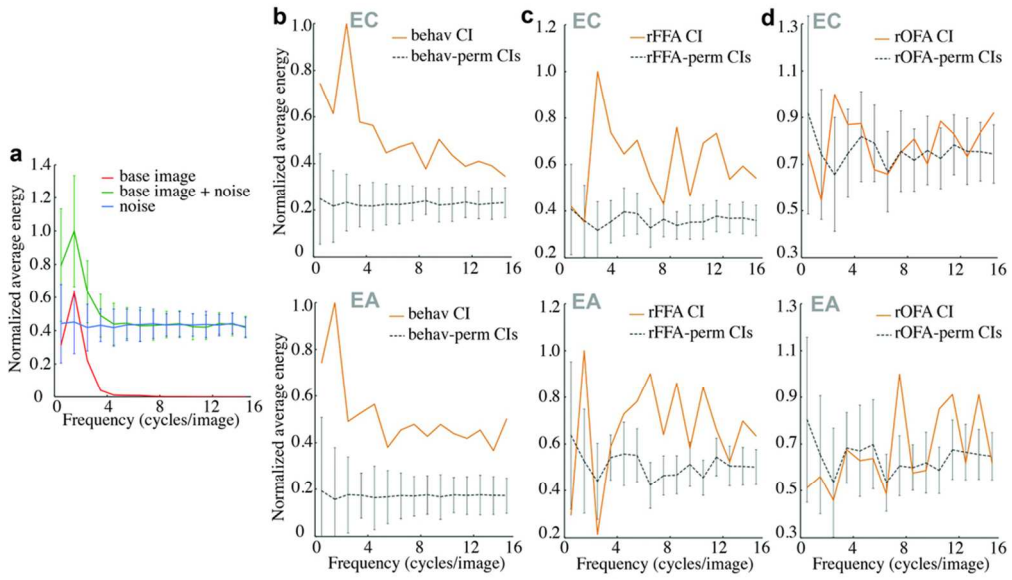


Fig 6. Average squared amplitude energy for (a) the base image, stimuli containing the base image and stimuli containing only noise fields; (b) the raw behavioral CIs; (c), (d) the raw neurally-derived CIs (corresponding to the right FFA and OFA). The abscissa represents spatial frequency in cycles per image and the ordinate displays normalized amplitude values averaged across orientations – values are normalized (scaled) by the maximum value. The average energy of 100 control CIs (constructed by permuting response labels) is shown in gray. Error bars show ±1SD across stimuli for (a) and across control CIs for (b-d).
94x54mm (300 x 300 DPI)

Review

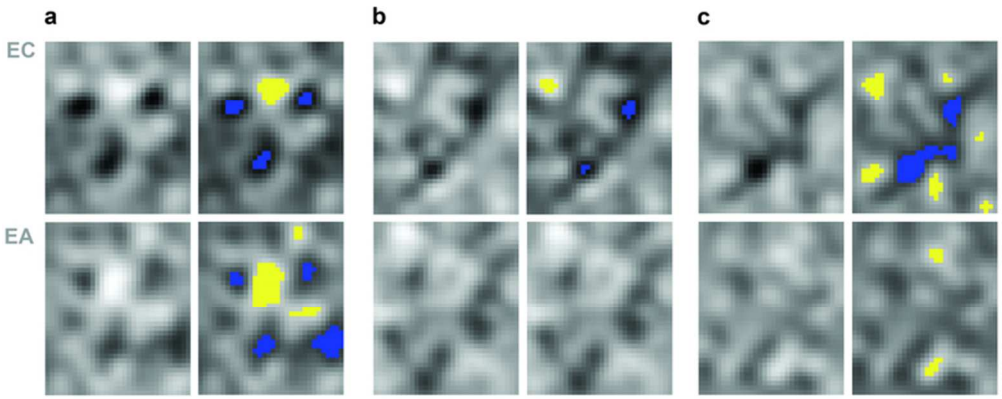


Fig 7. Smoothed CIs and their pixel test analysis. Results are shown for (a) behavioral responses, (b) right FFA responses and (c) the right OFA responses. Blue and yellow mark pixels darker / brighter than chance ($p<0.05$).
68x28mm (300 x 300 DPI)

Peer Review

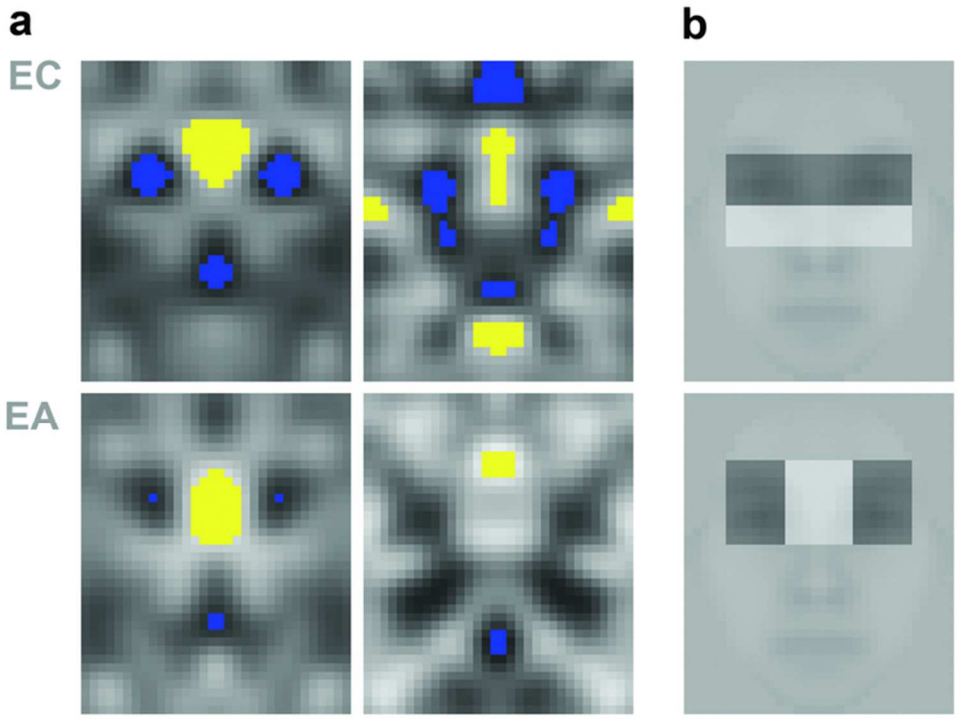


Fig 8. (a) Symmetrical CIs analyzed with a pixel test ($p<0.05$). Results are shown for behavioral CIs (on the left) and for rFFA-derived CIs (on the right). (b) The two best contrast features for face detection of Viola and Jones [2004] superimposed on a base image.
66x49mm (300 x 300 DPI)

eview