Systems/Circuits

Clutter Substantially Reduces Selectivity for Peripheral Faces in the Macaque Brain

Jessica Taubert,1,2 ⌂ Susan G. Wardle,1 ⌂ Clarissa T. Tardiff,1 Amanda Patterson,1 David Yu,3 and Chris I. Baker1

1Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, Maryland 20814, 2School of Psychology, The University of Queensland, Brisbane, Queensland 4072, Australia, and 3Neurophysiology Imaging Facility, National Institutes of Health, Bethesda, Maryland 20814

According to a prominent view in neuroscience, visual stimuli are coded by discrete cortical networks that respond preferentially to specific categories, such as faces or objects. However, it remains unclear how these category-selective networks respond when viewing conditions are cluttered, i.e., when there is more than one stimulus in the visual field. Here, we asked three questions: (1) Does clutter reduce the response and selectivity for faces as a function of retinal location? (2) Is the preferential response to faces uniform across the visual field? And (3) Does the ventral visual pathway encode information about the location of cluttered faces? We used fMRI to measure the response of the face-selective network in awake, fixating macaques (two female, five male). Across a series of four experiments, we manipulated the presence and absence of clutter, as well as the location of the faces relative to the fovea. We found that clutter reduces the response to peripheral faces. When presented in isolation, without clutter, the selectivity for faces is fairly uniform across the visual field, but, when clutter is present, there is a marked decrease in the selectivity for peripheral faces. We also found no evidence of a contralateral visual field bias when faces were presented in clutter. Nonetheless, multivariate analyses revealed that the location of cluttered faces could be decoded from the multivoxel response of the face-selective network. Collectively, these findings demonstrate that clutter blunts the selectivity of the face-selective network to peripheral faces, although information about their retinal location is retained.

Key words: clutter; face perception; face-selective; macaque fMRI; multiple objects; periphery

Significance Statement

Numerous studies that have measured brain activity in macaques have found visual regions that respond preferentially to faces. Although these regions are thought to be essential for social behavior, their responses have typically been measured while faces were presented in isolation, a situation atypical of the real world. How do these regions respond when faces are presented with other stimuli? We report that, when clutter is present, the preferential response to foveated faces is spared but preferential response to peripheral faces is reduced. Our results indicate that the presence of clutter changes the response of the face-selective network.

Received Jan. 31, 2022; revised Apr. 29, 2022; accepted June 6, 2022.

Author contributions: J.T., S.G.W., and C.I.B. designed research; J.T., C.T.T., A.P., and D.Y. performed research; J.T. and S.G.W. contributed unpublished reagents/analytic tools; J.T. and S.G.W. analyzed data; J.T. wrote the first draft of the paper; S.G.W. and C.I.B. edited the paper; J.T., S.G.W., and C.I.B. wrote the paper.

This work was supported by the Intramural Research Program of the National Institute of Mental Health (NIMH) Grants ZIAMH002918 (to Leslie G. Ungerleider) and ZIAMH002909 (to C.I.B.) and by the Australian Research Council Grant FT200100843 (to J.T.). Dr. Leslie G. Ungerleider supervised the early stages of this project (experiments 1 and 2); however, she passed away before the final results were available and could not assist with the analysis nor the design of follow-up experiments (experiments 3 and 4). Thus, the manuscript in no way reflects her interpretation of the data. We thank the Neurophysiology Imaging Facility Core (NIMH, National Institute of Neurological Disorders and Stroke, National Eye Institute) for functional and anatomical MRI scanning, with special thanks to Aidan Murphy, Charles Zhu, and Frank Ye for technical assistance.

The authors declare no competing financial interests.

Correspondence should be addressed to Jessica Taubert at j.taubert@uq.edu.au.

https://doi.org/10.1523/JNEUROSCI.0232-22.2022

Copyright © 2022 the authors

Introduction

Making sense of the world around us from the dense visual input that is collected on the retina is a computationally daunting task and one of the brain’s greatest accomplishments. The network of face-selective regions that has been identified in the primate visual system (Kanwisher et al., 1997; Tsao et al., 2003; Rossion et al., 2012; Hung et al., 2015) has proven to be a valuable model for understanding how the brain builds meaningful representations of visual stimuli (Orban et al., 2014; Hesse and Tsao, 2020), however, our understanding of this network is constrained by several experimental limitations. For example, most studies that have examined the responsibility of the face-selective network have done so by presenting isolated face stimuli at fixation (Kanwisher et al., 1997; Leopold et al., 2006; Freiwald and Tsao, 2010; Bell et al., 2011; Popivanov et al., 2012; Taubert et al., 2015, 2019).
2018b; Premereur et al., 2016; Wardle et al., 2020), yet this is a situation atypical of the real world where faces are present in clutter and potentially across the visual field. Therefore, what is needed is a better understanding of how the face-selective network operates under more natural viewing conditions (Leopold and Park, 2020; Wardle and Baker, 2020; Fan et al., 2021).

Previous studies have identified the face-selective network in the macaque brain (Tsao et al., 2003; Freiwald, 2020; Taubert et al., 2020a). It has been reported previously that when two stimuli are presented simultaneously with a fixation point in between, face-selective neurons continue to respond to the presence of a face (Zoccolan et al., 2005; Bao and Tsao, 2018). However, in these studies the stimuli were presented in peripheral vision, equidistant from a central fixation point. We would argue that, normally, when a face is present but not fixated, that is likely because the subject is focused on another object in the visual field. Indeed, when a stimulus occupies foveal vision, its representation may dominate the ventral visual pathway (Ishai et al., 1999; Haxby et al., 2001). Further, we know that the ventral visual pathway is strongly biased toward foveal inputs (Frisén and Glansholt, 1975). Thus, an open question is whether the face-selective network continues to respond preferentially to faces when another stimulus is being fixated.

Here, we ask three fundamental questions about the responsivity of face-selective network in the macaque brain. First, does clutter reduce the response to faces as a function of retinal location? To address this question, we devised a series of functional imaging (fMRI) experiments where we presented faces to awake macaques in one of three retinal locations (at fixation, 8° to the left, and 8° to the right) while nonface stimuli occupied the two other remaining retinal locations. Although previous studies have demonstrated that the responses of face-selective populations are impervious to clutter (Zoccolan et al., 2005; Reddy and Kanwisher, 2007; Agam et al., 2010; Bao and Tsao, 2018) in the current study we presented three-item horizontal displays and required the subject to fixate on the central item. Thus, our expectation was that clutter would decrease the response to peripheral faces, while the response to foveal faces would be spared (Rolls et al., 2003). Second, is the preferential response to faces, at the population level, uniform across the visual field? To address this question, we scanned subjects while they were presented with isolated faces and objects, while manipulating their retinal location. Given the size of receptive fields in the posterior face patch (Issa and DiCarlo, 2012), we expected face-responsivity to be lower in the peripheral locations compared with the central foveal location. Third, does the ventral visual pathway encode information about the location of cluttered faces? We reasoned that even if the magnitude of the response to faces was attenuated by clutter, activity in the face-selective network might still encode the location of faces. To explore this possibility, we examined the pattern of activity across voxels in response to face stimuli using multivariate analyses.

Materials and Methods

Subjects

We tested seven rhesus macaques (Macaca mulatta, 6–13 years of age, 7–12.1 kg at time of testing). Two were female (subjects S and R). Since previous reports of similar experiments on this species have demonstrated that sample sizes of two to four are sufficient for scientific inference (Hadj-Bouziane et al., 2008; Fisher and Freiwald, 2015; Russ and Leopold, 2015; Taubert et al., 2020a, 2022; Zhang et al., 2020) for each separate experiment we recruited two to four subjects. All subjects were acquired from the same primate breeding facility in the United States where they had social group histories as well as group housing experience until their transfer to the National Institute of Mental Health (NIMH) for quarantine approximately at the age of four years. After that, they were housed in a large colony room with auditory and visual contact with other conspecifics.

Subjects were surgically implanted with a headpost under sterile conditions using isoflurane anesthesia. For all subjects, the location of the headpost was planned to optimize the coverage of the inferior temporal lobe in fMRI experiments, at the expense of covering the parietal and frontal lobes. After recovery, the subjects were slowly acclimated to the experimental procedure; first they were trained to sit calmly in a restraint chair and fixate a small [0.4 degrees of visual angle (dva)] red central dot for long durations (~8 min). During acclimation and training, fixation within a circular window (radius = 2 dva) centered over the fixation dot resulted in juice delivery. All procedures were in accordance with the Guide for the Care and Use of Laboratory Animals and were approved by the NIMH Animal Care and Use Committee.

Data acquisition

Before each scanning session, an exogenous contrast agent [monocrystalline iron oxide nanocolloid (MION)] was injected into the femoral vein to increase the signal-to-noise ratio (Vanduffel et al., 2001; Taubert et al., 2020a). MION doses were determined independently for each subject (~8–10 mg/kg).

Structural and functional data were acquired in a 4.7T, 60 cm vertical scanner (Bruker Biospec) equipped with a Bruker 5380 gradient coil. Subjects viewed the visual stimuli projected onto a screen above their head through a mirror positioned in front of their eyes. We collected whole brain images with a four-channel transmit and receive radio frequency coil system (Rapid MR International). A low-resolution anatomic scan was also acquired in the same session to serve as an anatomic reference [modified driven equilibrium Fourier transform (MDEFT) sequence, voxel size: 1.5 × 0.5 × 0.5 mm, FOV: 96 × 48 mm, matrix size: 192 × 96, echo time (TE): 3.95 ms, repetition time (TR): 11.25 ms]. Functional echoplanar imaging (EPI) scans were collected as 42 sagittal slices with an in-plane resolution of 1.5 × 1.5 mm and a slice thickness of 1.5 mm. The TR was 2.2 s and the TE was 16 ms (FOV: 96 × 54 mm, matrix size: 64 × 36 m: flip angle 75°). Eye position was recorded using an MR-compatible infrared camera (MRC Systems) fed into MATLAB (MathWorks, version R2018b) via a DATAPixx hub (VPixx Technologies, Vision Science Solutions).

Localization data

First, we identified the regions of inferior temporal cortex (ITC) that responded preferentially to faces in all seven subjects using a standard face localization procedure (Tsao et al., 2003; Premereur et al., 2016; Taubert et al., 2020a, b). While the subjects were awake and fixating, we presented images (30 per category) of six different object categories (human faces, monkey faces, scenes, objects, phase scrambled human faces and phase scrambled monkey faces). Stimuli were cropped images presented on a square canvas that was 12 dva in height. All six categories were presented in each run in a standard on/off block design (12 blocks in total). Each block lasted for 16.5 s. During a “stimulus on” block, 15 images were presented one at a time for 900 ms and were followed by a 200-ms interstimulus interval (ISI). We removed any run from the analysis where the monkey did not fixate within a 4° window for >60% of the time.

Face-selective voxels in ITC were identified in all seven subjects using the following contrast: activations evoked by (human faces + monkey faces) > activations evoked by (scenes + objects; Fig. 1A). For every subject we used the same statistical threshold \( p = 0.001 \); controlled for multiple comparisons using the false discovery rate (FDR)); this yielded a different number of voxels for each subject [experiment 1: subject F = 577 voxels (\( t = 309, r = 268 \)); subject K = 426 voxels (\( t = 197, r = 229 \)); experiments 2–4: subject A = 443 voxels (\( t = 236, r = 207 \)); subject H = 393 voxels (\( t = 331, r = 262 \)); subject J = 523 voxels (\( t = 253, r = 270 \)); subject S = 773 voxels (\( t = 387, r = 386 \)); subject R = 519 voxels (\( t = 279, r = 240 \))]. In each of the seven subjects, this procedure identified all of the core face-selective regions that have been
we subtract the response to a scrambled face condition from the scrambled face conditions was the face stimulus used. Thus, when each triplet were allocated to left or right locations at random. The only levels of "face stimulus at fixation" were isolated at fixation, cluttered at fixation, cluttered left; and cluttered right; (Fig. 4C). We controlled the stimulus that was pre-published. We included face and object isolated conditions to verify that the new stimulus set was comparable to the one used for experiment 2 (Fig. 4C). Block duration was 17.6 s in experiment 3, and every run was 286 s in length.

**Experiment 4**

Stimuli were presented either alone at any one of the three visual field locations, or in triplets (Fig. 5A). The stimuli were the 15 faces, 15 objects, and 15 scenes used in experiment 1. The repeated measures design included six conditions; two levels of clutter (isolated or cluttered) and three levels of retinal location (at fixation, left, right; Fig. 5A). All timing parameters were the same as those described for experiment 1.

**Fixation behavior**

During every experiment, subjects were rewarded for maintained fixation at random intervals. The average time between rewards was typically 2 s (± 400 ms) but these parameters were modified depending on the subject's behavior. Subjects had to fixate for at least 85% of the total run length (excluding the initial fixation period) for any given run to be included in the final analysis. We noted, however, that while the subjects were generally compliant during stimulation blocks, they tended to rest during the fixation blocks by either looking down or briefly closing their eyes. Since resting during fixation blocks was not problematic for the analysis, we placed a second behavioral criteria on each run; subjects also had to fixate for at least 95% of the total length of the condition blocks (concatenated).

**Experiment 1**

Stimuli were shown in triplets, with one stimulus at central fixation, flanked by a peripheral stimulus on each side (Fig. 1B). The stimuli were 15 macaque faces, 15 inanimate objects, 15 scenes, and 15 scrambled macaque faces. We paired each of the 15 macaque faces with a particular object and scene, creating 15 stimulus triplets in total. Using these triplets, we devised 6 unique conditions in a $2 \times 3$ repeated measures design; two levels of facial structure (intact or scrambled) and three levels of the "stimulus at fixation" manipulation (face, object, or scene). Across the levels of "stimulus at fixation" we controlled the stimulus that was presented in the central fixation location. The remaining two stimuli in each triplet were allocated to left or right locations at random. The only difference between the trials in the intact face conditions and those in the scrambled face conditions was the face stimulus used. Thus, when we subtract the response to a scrambled face condition from the corresponding intact face condition, the only difference between those conditions was the presence of a face. The specific object and scene were the same, and the relative location of all three stimuli was also the same. In experiment 1, every block lasted for 16.5 s. The order of the conditions, and the triplets within each condition, was determined at random at the beginning of each run. Each run lasted 202.4 s during which we collected 92 volumes of data.

**Experiment 2**

In this experiment, only one stimulus was on screen at a time (Fig. 2A). The stimuli were photographs of 15 macaque faces and 15 objects. These were not the same stimuli used in experiment 1. The repeated measures design included six conditions; two levels of stimulus (faces or objects) and three levels of retinal location (at fixation, left, right; Fig. 2A and Fig. 3). All timing parameters were the same as those described for experiment 1.

**Experiment 3**

The stimuli were presented either alone at central fixation, or in triplets (Fig. 4C). The stimuli were 16 macaque faces, 16 objects, and 32 photographs of 15 macaque faces and 15 objects. These were not the same stimuli used in experiment 1 or 2. Each laboratory scene was paired with a natural scene. These 16 unique pairs of scenes were each assigned to a particular object and a particular face (for an illustrative example, see Fig. 4C). In this factorial design, there were two levels of stimulus (face or object) and four levels of presentation (isolated at fixation, cluttered at fixation, cluttered left and cluttered right; Fig. 4C). We included face and object isolated conditions to verify that the new stimulus set was comparable to the one used for experiment 2 (Fig. 4C). Block duration was 17.6 s in experiment 3, and every run was 286 s in length.

**Figure 1.** Reduced responses to peripheral faces in the presence of clutter. A, The face-selective network identified in subject H's data using a statistical threshold of $q = 0.0001$ False Discovery Rate (FDR). B, Illustrative examples of the stimulus arrays used across the six conditions in experiment 1. C, Face responsivity scores for all face-selective voxels as a function of the stimulus at fixation manipulation. The median difference scores were: faces at fixation $= 0.39$, objects at fixation $= 0.03$, scenes at fixation $= 0.12$.

**Visual stimulation**

Four different fMRI experiments were conducted. In every experiment we presented visual stimuli in on/off block-designs. Every run began with two dummy pulses, and then 4.4 s of fixation before the onset of the experiment. This brief fixation period was included to help get the subject settled and was excluded from further analysis. Experimental runs included a number of stimulation blocks that were each followed by a fixation block of equal length. The exact number of stimulation blocks differed across experiments, always matching the number of experimental conditions. During stimulation blocks we presented stimulus displays containing single items or three items. Each stimulus display was presented for 900 ms, followed by a 200-ms ISI.

All stimuli (faces, objects, scenes, scrambled faces) were cropped to a square shape (6 dva in height) but color information was preserved. Any given stimulus was presented in one of three horizontal screen locations: at fixation, left of fixation and right of fixation. The center of a stimulus at fixation was the exact center of the screen (0,0). The center of a stimulus presented left of fixation was 8° away from the center of the screen (−8.0). Similarly, the center of a stimulus presented right of fixation was also 8° away from the center of the screen (+8.0).

Experiment 1

Stimuli were shown in triplets, with one stimulus at central fixation, flanked by a peripheral stimulus on each side (Fig. 1B). The stimuli were 15 macaque faces, 15 inanimate objects, 15 scenes, and 15 scrambled macaque faces. We paired each of the 15 macaque faces with a particular object and scene, creating 15 stimulus triplets in total. Using these triplets, we devised 6 unique conditions in a $2 \times 3$ repeated measures design; two levels of facial structure (intact or scrambled) and three levels of the "stimulus at fixation" manipulation (face, object, or scene). Across the levels of "stimulus at fixation" we controlled the stimulus that was presented in the central fixation location. The remaining two stimuli in each triplet were allocated to left or right locations at random. The only difference between the trials in the intact face conditions and those in the scrambled face conditions was the face stimulus used. Thus, when we subtract the response to a scrambled face condition from the...
and subject R completed 11 runs in just one session (1408 volumes). Finally, four subjects participated in experiment 4 (A, H, S, and R; two female). After applying the two-step behavioral criteria to remove poor quality runs, subject A completed 24 runs in one scan session (2160 volumes); subject H completed 40 runs (3600 volumes) across three sessions; subject S completed 15 runs in just one session (1350 volumes), and subject R completed 20 runs (1800 volumes) across two sessions.

We analyzed the eye-tracking data in experiments 3 and 4 to confirm that fixation behavior was similar across clutter absent and clutter present conditions. To do this we computed the proportion of time spent fixating for each unique condition (i.e., block) in every valid run that was included in the fMRI analyses. In experiment 3, the subjects completed 64 valid runs in total. We compared fixation behavior across the eight unique conditions in a one-way analysis of variance for repeated measures and found no evidence that fixation behavior varied across conditions ($F_{(7,441)} = 0.65, p = 0.71, \eta^2_p = 0.01$). In experiment 4, the subjects completed 99 valid runs in total. We compared fixation behavior across the six unique conditions in a one-way analysis of variance for repeated measures and found no evidence that fixation behavior varied across conditions ($F_{(5,490)} = 0.37, p = 0.87, \eta^2_p = 0.004$).

**fMRI data analysis**

To facilitate cortical surface alignments, we acquired high-resolution T1-weighted whole-brain anatomic scans in a 4.7T Bruker scanner with an MDEFT sequence. Imaging parameters were as follows: voxel size: $0.5 \times 0.5 \times 0.5$ mm, TE: 4.9 ms, TR: 13.6 ms, flip angle: 14'. All EPI data were analyzed using AFNI software (Cox, 1996; http://afni.nimh.nih.gov/afni). Raw images were first converted from Bruker into AFNI data file format. The data collected in a single session were first corrected for static magnetic field inhomogeneities using the PLACE algorithm (Xiang and Ye, 2007). The time series data were then slice-time corrected and realigned to the last volume of the last run. All the data for a given subject were registered to the corresponding high-resolution anatomic scan of that subject, allowing for the combination of data across multiple sessions. Thus, all data were analyzed in individual subject space. The first two volumes of data in each EPI sequence were disregarded. The volume registered data were then despiked and spatially resampled with a $3\text{ mm}$ Gaussian kernel and rescaled to reflect percent signal change from baseline.

We convolved the hemodynamic response function for MION exposure with the regressors of interest using an ordinary least squares regression (executed using the AFNI function ‘3dDeconvolve’ with ‘MIONN’ as the response function). The regressors of no interest included in the model were 6 motion regressors (movement parameters obtained from the volume registration) and AFNI’s baseline estimates and signal drifts (linear and quadratic).

When comparing $\beta$ coefficients, we wanted to ensure that results were not influenced by small numbers of voxels with extreme $\beta$ coefficients and, thus, we normalized the data using the min-max method within each subject ($z_i = (x_i - \min(s))/(\max(s) - \min(s))$). All statistical comparisons were performed using custom scripts written in MATLAB (MathWorks, version R2018b).

**Multivariate pattern analyses (experiment 4)**

In experiment 4, we used a whole-brain decoding searchlight analysis which was implemented using The Decoding Toolbox (TDT; Hebart et al., 2014). The decoding was performed using a Newton linear SVM classifier with a radius of 3 voxels in each subject’s native space. For cross-validation we used a leave-one-run-out procedure. Decoding accuracy was averaged across all cross-validation folds. To locate the areas that decoded face location when clutter was absent (C–), we performed three pairwise classifications; (1) Fixation_C– versus Left_C–, (2) Fixation_C– versus Right_C–, and (3) Left_C– versus Right_C–. Then we averaged decoding accuracy across all three classifications. Similarly, to locate the areas that decoded face location when clutter was present (C+), we performed three pairwise classifications; (1) Fixation_C+ versus Left_C+ (2) Fixation_C+ versus Right_C+ and (3) Left_C+ versus Right_C+. Then we averaged decoding accuracy across all three classifications. For this analysis, numerical chance was 50%.

**Results**

**Loss of responsivity to peripheral faces when clutter is present (experiment 1)**

Our first experiment was designed to measure the response of the face-selective network (Fig. 1A) to faces when they were...
presented in three-item cluttered displays. In every condition, a face (scrambled or intact) was presented alongside an object and a scene. Across conditions, we manipulated the identity of the stimulus that was presented at fixation (faces, objects, or scenes; Fig. 1B). We measured “face-responsivity,” defined as the response to intact faces relative to phase-scrambled faces (effectively noise but with similar low-level visual statistics to actual faces) in the same configuration. Thus, for every face-selective voxel we calculated the difference between the response to intact and scrambled faces (i.e., face-responsivity = β [intact faces] − β [scrambled faces]). In Figure 1C, we plot face-responsivity as a function of the stimulus at fixation. Overall, we found that face-responsivity was statistically higher when faces were being fixated than when objects (N = 1003, Z = −17.75, p < 0.0001; Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons) or scenes (N = 1003, Z = −14.67, p < 0.0001) were fixated.

To show that these findings replicated across subjects, we performed the same analysis separately for each subject. For subject K, we calculated face-responsivity across the three conditions (faces, objects, or scenes) for every voxel (x faces = 0.33; x objects = 0.10; x scenes = −0.009). Using related-samples Wilcoxon signed-rank tests, we found that face-responsivity was statistically higher when faces were being fixated than when objects (N = 426, Z = −7.07, p < 0.0001) or scenes (N = 426, Z = −8.87, p < 0.0001) were fixated. When calculating face-responsivity for subject F (x faces = 0.44; x objects = −0.01; x scenes = 0.21), we found evidence of the same differences between conditions (faces > objects, N = 577, Z = −17.51, p < 0.0001; faces > scenes, N = 577, Z = −11.71, p < 0.0001). Therefore, our findings replicated across subjects.

These findings suggest that the face-selective network responds less to peripheral faces than foveal faces when clutter is present. However, it is not clear from these results whether the reduction in the response to face stimuli was driven by the presence of clutter per se or simply by the use of peripheral screen locations. Thus, we next measured the response of the face-selective voxels to isolated faces and objects presented in the same three screen locations used in experiment 1 to determine whether the face-selective network responds preferentially to peripheral faces compared with other stimuli, when clutter is not present.

The face-selective network responds more to peripheral faces than peripheral objects (experiment 2)
The goal of experiment 2 was to determine whether activity in the putative face-selective network is driven more by peripheral faces than peripheral objects in the absence of clutter. To this end, we ran a block-design fMRI experiment with six conditions, manipulating both the visual field location of the single stimulus presented on each trial (at fixation vs left hemifield vs right hemifield) and the stimulus category (faces vs objects; Fig. 2A). We included both faces and objects in this design so that we could measure face-selectivity (i.e., the preferential response to faces over objects) rather than face-responsivity. For the purposes of this analysis, we computed the responses in terms of the contra-lateral and ipsilateral responses (Fig. 2B), and we normalized the data within subject using the min-max method.

Using a set of three related samples Wilcoxon signed-rank tests we found that the response to isolated faces was maximal when presented at fixation (fixation vs contralateral, N = 1366, Z = −22.94, p < 0.0001; fixation vs ipsilateral, N = 1366, Z = −31.83, p < 0.0001; Fig. 2C). These results are consistent with the overrepresentation of foveal vision in temporal cortex (Dow et al., 1981; Van Essen et al., 1984). Further, we also found evidence that the magnitude of the response to faces in the contra-lateral hemifield was greater than the magnitude of the response to faces in the ipsilateral hemifield (N = 1366, Z = −28.34, p < 0.0001; Fig. 2C). The observed contralateral bias is also consistent with the current models of visuospatial encoding in the ventral visual pathway (Silson et al., 2021; Groen et al., 2022).

Next, we measured the preferential response to faces by computing a selectivity index for each voxel using the following equation: face selectivity index (FSI) = [β faces − β objects]/[|β faces| + |β objects|] (Tsao et al., 2006; Taubert et al., 2020a). We used three one-sample Wilcoxon signed-rank tests to test the null hypothesis that the median FSI value was 0 (i.e., there was no preference toward faces over objects in any of the location conditions). In doing so, we discovered that face-selective voxels continue to respond more to faces than objects, regardless of their location (fixation, median = 0.18, N = 1366, Z = 25, p < 0.0001; contralateral, median = 0.16, N = 1366, Z = 21.83, p < 0.0001; ipsilateral, median = 0.14, N = 1366, Z = 16.77, p < 0.0001). Nonetheless, FSI values were higher in the foveal condition than the two peripheral conditions (fixation vs contralateral, N = 1366, Z = −4.27, p < 0.0001; fixation vs ipsilateral, N = 1366, Z = −2.94, p = 0.003; Fig. 2D). Further, there was no evidence of a contralateral bias (contralateral vs ipsilateral, N = 1366, Z = −1, p = 0.31; Fig. 2D).

Collectively, these findings indicate that, while the selectivity for face stimuli is not abolished by peripheral shifts of 8 dva, it is slightly reduced.

For subject S, we confirmed that the face-selective voxels responded more to faces than objects, regardless of their screen location (fixation, median = 0.10, N = 773, Z = 14.28, p < 0.0001; contralateral, median = 0.08, N = 773, Z = 11.03, p < 0.0001; ipsilateral, median = 0.09, N = 773, Z = 8.46, p < 0.0001). This was also true for subject F (fixation, median = 0.27, N = 593, Z = 20.03, p < 0.0001; contralateral, median = 0.22, N = 593, Z = 19.13, p < 0.0001; ipsilateral, median = 0.23, N = 593, Z = 14.98, p < 0.0001).

To validate the region-of-interest approach, we visualized the whole-brain contrast between faces and objects separately as a function of visual field location in the anatomic volume for individual subjects (Fig. 3A). These whole-brain contrasts indicated that the location of the face-selective voxels in ITC was stable across the three location conditions. We also examined the strength of the relationship between the three face conditions (fixation, left, and right conditions). We found that there were significant positive relationships between all pairs of conditions (fixation vs left, Spearman’s ρ(1366) = 0.9, p < 0.0001; fixation vs right, Spearman’s ρ(1366) = 0.83, p < 0.0001; left vs right, Spearman’s ρ(1366) = 0.71, p < 0.0001). Thus, we are confident that the reductions in selectivity that we observed for peripheral faces were not driven by the fact that presenting faces in different retinotopic locations recruits different neural populations. In sum, the results of experiment 2 confirm that the face-selective network responds more to peripheral faces than peripheral objects when clutter is not present. In the next experiment, we used the same experimental design, except we added visual clutter to all stimulus displays. We compare the results to those of experiment 2 to determine whether the tolerance of peripheral shifts survives the addition of clutter.

The face-selective network responds less selectively to peripheral faces when clutter is present (experiment 3)
In experiment 3, our goal was to measure the selectivity for face stimuli across the visual field, as in experiment 2, but in the presence of clutter. We designed an experiment with eight conditions...
in which we manipulated the relative location of faces and objects in three-item cluttered displays (Fig. 4C). To determine whether face-selectivity is impervious to clutter, we compared face-selectivity when faces/objects were presented at fixation, to face-selectivity when faces/objects were presented away from fixation, either in the contralateral or ipsilateral hemifield. We performed these analyses, separately, for each subject.

For subject S, we found that face-selectivity (i.e., FSI values) was significantly higher when faces/objects were presented at fixation compared with when they were presented in the contralateral (N = 773, Z = −5.37, p < 0.001) or ipsilateral hemifield (N = 773, Z = −15.17, p < 0.0001). We found the same pattern of results for subject R (fixation vs contralateral, N = 519, Z = −18.63, p < 0.0001; fixation vs ipsilateral, N = 519, Z = −9.09, p < 0.0001) and subject J (fixation vs contralateral, N = 523, Z = −19.14, p < 0.0001; fixation vs ipsilateral, N = 523, Z = −18.88, p < 0.0001). These results indicate that selectivity is reduced for peripheral faces, compared with foveal faces, when clutter is present. As such, these results appear at odds the results of experiment 2. Next, we compared the results of experiments 2 and 3 more directly.

Clutter reduces the selectivity for peripheral faces (comparison of experiments 2 and 3)

Since experiments 2 and 3 employed different stimuli, we used the isolated face and object conditions, which were present in both experiments, to determine whether the stimulus sets evoked the same response from the face-selective voxels. After normalizing the data within subject, we used two Mann–Whitney U tests (two-tailed, controlled for multiple comparisons using the Bonferroni rule) to test for differences in the normalized fMRI signal across the two experiments. We found no evidence that the face stimuli elicited a differential response between experiments (N = 3181, Z = −0.27, p = 0.78; Fig. 4B). Similarly, we found no evidence that the object stimuli elicited a differential response between experiments (N = 3181, Z = −0.09, p = 0.93; Fig. 4B). Therefore, we moved forward with a comparison between experiments 2 and 3.

To determine whether the preferential response to faces over objects was impacted by the addition of clutter, we used a set of three Mann–Whitney U tests across experiments 2 and 3 (two-tailed, controlled for multiple comparisons using the Bonferroni rule).
Again, the only difference between the fixation, contralateral and ipsilateral conditions across the two experiments was that in experiment 2 the stimuli were presented in isolation, whereas in experiment 3 the stimuli were presented in clutter. In the fixation condition, we discovered that the distribution of FSI values measured in experiment 2 (median FSI = 0.18; Fig. 2D) was different from the distribution of FSI values measured in experiment 3 (median FSI = 0.18, N = 3181, Z = 1.25, p = 0.21; Fig. 4D) indicating that, when faces and objects are being foveated, the preferential response to faces is impervious to clutter.

In contrast, when peripheral faces were presented in the contralateral hemifield, we found that the FSI values measured in experiment 2 (median FSI = 0.16; Fig. 2D) were much higher than those measured in experiment 3 (median FSI = 0.02, N = 3181, Z = −18.64, p < 0.0001; Fig. 4D). This was also true when peripheral faces were presented in the ipsilateral hemifield (median FSI (experiment 2) = 0.14 (Fig. 2D); median FSI (experiment 3) = −0.0007, N = 3181, Z = −16.94, p < 0.0001 (Fig. 4D)). These results indicate that the preference for peripheral, but not foveal, faces is significantly reduced by clutter.

However, because the observations in experiments 2 and 3 were based on a different number of subjects, and each subject completed a different number of runs, we could not compare the isolated and cluttered conditions using a more powerful within-subjects analysis. This approach also prevented the employment of multivariate analyses to examine the pattern of activity across face-selective voxels. Thus, the final experiment (experiment 4) was designed so that brain activity in response to isolate faces and cluttered faces could be more directly compared within subjects.

**Does selectivity for peripheral faces increase toward the anterior pole?**

It has been suggested that the size of neuronal receptive fields may systematically increase toward the anterior regions of the ventral visual pathway (Desimone and Gross, 1979; Ito et al., 1995; also see Silson et al., 2021). Thus, it is possible that the anterior regions of the face-selective network might respond more selectively for peripheral face stimuli than the posterior regions. To test this idea, we examined the relationship between a voxel’s relative location in the brain, along the posterior-anterior (P-A) axis, and its FSI across our experimental conditions (experiment 2, fixation, contralateral, ipsilateral; experiment 3, fixation, contralateral, ipsilateral).

To do this we took the P-A coordinates of all the face-selective voxels in an individual subject’s mask and standardized them with reference to the most posterior voxel (i.e., P-A coordinates; 0 = the most posterior face-selective voxels, >0 = relative distance, in 1.5-mm increments, from most anterior face-selective voxels). Next, we used Spearman correlations to assess the relationship between relative anatomic location (relative P-A coordinates; Fig. 5) and FSI, expecting negative relationships because single unit recordings have indicated that the posterior regions are generally more face-selective than the anterior regions for foveal stimuli, perhaps owning to changes in how discrete identities are represented (Freiwald and Tsao, 2010; Bell et al., 2011; Taubert et al., 2015, 2018b). The results revealed significant negative relationships across all conditions in experiment 2, when clutter was absent (Fig. 5). This was also true when clutter was present and the face stimuli were presented away from fixation (Fig. 5); however, when clutter was present and the face stimuli were presented at fixation the direction of the relationship between relative anatomic location and FSI was altered (Fig. 5). Fishers z-transformations were used to compare the correlations across the clutter absent and clutter present conditions. This analysis revealed evidence that, when faces were presented at fixation, there was a stronger correlation when clutter was present than when clutter was absent with a change in the sign of the correlation (z = −12.74, p < 0.001, two-tailed, the observed p-value has been corrected for multiple comparisons using the Bonferroni rule). Similarly, when faces were presented in the contralateral hemifield there was a stronger negative correlation when clutter was present than when clutter was absent (z = 2.45, p = 0.03, two-tailed, p-value corrected for multiple comparisons). In contrast, when faces were presented in the ipsilateral hemifield, we found no evidence that the negative correlations differed (z = −2.04, p = 0.12, two-tailed, p-value corrected for multiple comparisons).

**Three-item displays evoke more activity than isolated faces (experiment 4)**

In the fourth and final experiment we employed a powerful repeated measures design to examine the response to faces in the three retinal locations while also manipulating the presence of clutter (Fig. 6A). First, we subtracted the normalized response to the cluttered condition from the corresponding isolated condition for each of the three levels of location (Fig. 6B). This analysis revealed that, in general, the face-selective voxels responded more to the cluttered conditions than the isolated conditions [median diff (fixation) = −0.06, null hypothesis is that the
The contralateral bias is eliminated by clutter (experiment 4)

We computed the contralateral bias for every face-selective voxel by subtracting the response to faces presented on the ipsilateral hemifield from the response to faces presented on the contralateral hemifield (i.e., clutter absent) and the cluttered face conditions. This revealed a contralateral bias in the isolated face conditions, when clutter was absent (N = 2328, Z = 37.1, p < 0.0001; Fig. 6C). In contrast, when clutter was present, there was evidence of a slight ipsilateral bias (N = 2328, Z = −3.33, p = 0.001; Fig. 6C).

To determine whether this finding was reliable across subjects, we compared the contralateral bias across clutter conditions, for each of the four subjects, separately. Consistent with the overall finding, we found that every subject had a significantly larger contralateral bias in the clutter absent condition than in the clutter present condition [subject H, (x absent = 0.04) > (x present = 0.01), N = 593, Z = −8.02, p < 0.001; subject A, (x absent = 0.1) > (x present = 0.02), N = 443, Z = −15.7, p < 0.0001; subject S, (x absent = 0.07) > (x present = −0.03), N = 773, Z = −22.26, p < 0.0001; subject R, (x absent = 0.09) > (x present = 0.02), N = 519, Z = −7.29, p < 0.001].

One interpretation of these results is that when there are multiple items present in the visual field, as there often are in complex natural scenes, the magnitude of response of the face-selective network does not reflect the horizontal location of a face. This raises the question, does the ventral visual pathway encode information about the location of visual stimuli at all or, alternatively, does activity in the ventral visual pathway primarily reflect only the form and identity of the foveated stimulus (Ishai et al., 1999; Haxby et al., 2001)?

Information about face location can be decoded from ITC (experiment 4)

Thus far, the results of this study (experiments 1–4) have repeatedly demonstrated that when clutter is present, and a nonface stimulus is being foveated, there is a reliable reduction in the response of face-selective voxels to peripheral faces (i.e., face responsivity and selectivity were significantly reduced, and the contralateral bias was eliminated). Yet, behavioral observations have indicated that human and nonhuman primates detect and orient their gaze toward faces that are initially presented away from fixation, prioritizing them over other targets and task objectives (Crouzet et al., 2010; Landman et al., 2014; Sadagopan et al., 2017; Taubert et al., 2017, 2018a; Keys et al., 2021). How are peripheral faces detected?

To determine whether the face-selective network encodes information about the location of peripheral faces we employed multivariate analyses to examine the data from experiment 4 (n = 4). We performed two decoding searchlight analyses in each subject’s native space to compare decoding of the face’s location in the visual field (at fixation, left periphery, or right periphery) separately for the clutter absent and the clutter present conditions (Fig. 7). These analyses revealed that when faces were presented alone (in the clutter absent conditions), the regions with the highest location decoding performance included early visual cortex (EVC). This is not surprising given the retinotopic nature of stimulus-evoked activity in early visual areas and strong contralateral biases that are known to influence activity in early visual areas (Silson et al., 2021; Groen et al., 2022). We note that successful location decoding is also possible further along the ventral visual pathway, which is consistent with the contralateral bias reported in experiment 2 (Fig. 7).

In contrast, when clutter was present, the results of the searchlight analysis changed; most notably the regions with the highest location decoding performance no longer included EVC.
Instead, we observed the highest decoding performance for the visual field position of faces in regions of the ventral visual pathway seated on the lower bank of the STS, the fundus of the STS, and in the lateral convexity (i.e., locations approximating the face-selective network). In Figure 8A, we compare the searchlight results across EVC and face-selective cortex (in the ventral visual pathway) and after combining the data across subjects. For voxels in the EVC, we found a decrease in classifier performance following the addition of visual clutter ($\bar{\text{perf}}_{\text{absent}} = 80.42\% > \bar{\text{perf}}_{\text{present}} = 57.08\%$, $N = 7560$, $Z = -74.74$, $p < 0.001$; Fig. 8A). However, for voxels in the face-selective network, when we compared classifier performance in the clutter absent
(\(\bar{x} = 64.44\%\)) and clutter present (\(\bar{x} = 65.55\%\)) conditions we found evidence that the classifier performed reliably better following the addition of visual clutter (absent < present, \(N = 2329, Z = 2.74, p = 0.006;\) Fig. 8A). These results confirm the observations in Figure 7. Out of interest, we also investigated the relationship between a face-selective voxel’s relative anatomic location and classifier performance. The results show that the posterior regions of face-selective network typically performed better than anterior regions of the face-selective network (Fig. 8B), but this is true across both clutter conditions (Fig. 8B).

Overall, these results provide some indication that the face-selective regions of the ventral visual pathway retain information about the location of faces in three item displays, even in the presence of clutter.

Discussion

Our motivation for this present study was to characterize how the putative face-selective network in the macaque brain responds to faces under cluttered viewing conditions. Across a series of four experiments, we showed that the face-selective network does respond preferentially to peripheral faces, relative to peripheral objects, however, this preferential response is quenched when another stimulus occupies foveal vision. In experiment 1, we found that face-selective voxels in ITC were less responsive to peripheral faces in clutter (Fig. 1C), and in experiment 3, we found that face-selective voxels in ITC exhibited reduced selectivity to peripheral faces in clutter (Fig. 4C).

Taken together, these findings provide much needed evidence that when faces are present in the periphery, but something else is being viewed, there is a significant reduction in the characteristic response profile of the face-selective network. Therefore, the tolerance of visual clutter that the face-selective network has exhibited in the past (Zoccolan et al., 2005; Reddy and Kanwisher, 2007; Bao and Tsao, 2018) is dependent on context.

These findings are consistent with the previous observation that ITC neurons have receptive fields that shrink in size following the addition of visual clutter (Rolls et al., 2003). It was asserted that this was a physiological response enabling ITC to provide an unambiguous representation of the stimulus at fixation. Indeed, the ventral visual pathway is optimized for processing foveal inputs (Frisén and Glansholm, 1975) and representing the fine-grained details of the faces and objects we are looking directly at, to the exclusion of all else (Ishai et al., 1999; Haxby et al., 2001; Rolls et al., 2003). It follows that, when clutter is present, the ventral visual pathway will prioritize the foveated stimulus and lose information about peripheral stimuli. Although, the results of experiment 4 reveal that the classically defined face-selective network retains information about the location of peripheral faces even when clutter is present (Fig. 7). Where this information is inherited from remains an open question that will need to be addressed by future research.

Interestingly, when we investigated the relationship between face selectivity and anatomic location in the brain, we found what we expected, face selectivity was higher in the posterior regions of the network than in the anterior regions of the network, except for one condition (Fig. 5). When foveated faces were presented in clutter, we found that FSI was greater in the anterior regions than in the posterior regions (Fig. 5). These observations further indicate that the presence of clutter changes how the face-selective network is operationalized, although exactly how increases and decreases in face-selectivity facilitate processing remains poorly understood. Again, these results highlight important gaps in our knowledge regarding how faces and other objects are processed by the visual system under more naturalistic demands.

A key feature of the current study is that we were able to distinguish between foveated and peripheral stimuli in clutter by using triplet displays with a single stimulus at fixation and two equidistant peripheral stimuli. Previous human fMRI studies investigating the representation of multiple objects have used two-item displays, with the object pairs equidistant from a central fixation point. These studies have reported that the responses to these two-item displays are well predicted by an averaging of the responses to the same objects presented in isolation (Macevoy and Epstein, 2009; Reddy et al., 2009; MacEvoy and Epstein, 2011; Baeck et al., 2013; Song et al., 2013). Further, in some areas of the ventral stream, the relationship between brain activity for isolated versus multiple objects is modulated by whether the object pairs are presented in a meaningful spatial configuration, such as a bottle positioned over a glass (Baeck et al., 2013; Quek and Peelen, 2020) or a person interacting with an object (Baldassano et al., 2017). The present study demonstrates that in addition to effects of spatial context, the response of the ventral visual pathway to multiple objects is also modulated by which item is fixated.

An outstanding question is how the visual system codes for the details of simultaneously viewed faces and objects. In the present study we found a marked reduction in the face-selective response when a nonface stimulus was being foveated. Previous studies using two-item peripheral displays reported that the category-selective response is preserved under conditions of clutter (Reddy and Kanwisher, 2007; Reddy et al., 2009), suggesting that the distinction between foveated and peripheral stimuli is important for understanding the coding of multiple objects. There is evidence from human intracranial recordings in temporal cortex for “robustness” to clutter even in the early part of the response profile (Agam et al., 2010). Robustness to clutter has been argued to be computationally advantageous for processing multiple stimuli (Cox and Riesenhuber, 2015), and the linear mixing of stimulus representations has been calculated to entail a significant cost to encoding accuracy (Orhan and Ma, 2015). Attention to a particular stimulus in multi-object displays modulates its...
representation (Kastner et al., 1998; Reddy et al., 2009), and thus attention may make an important contribution to the untangling of competing stimulus representations (Orhan and Ma, 2015).

Together, these results highlight the importance of increasing ecological validity when studying the visual system. We will only discover how vision is accomplished by the primate brain by placing realistic demands on the visual system (Leopold and Park, 2020; Fan et al., 2021). For example, in the present study we found no contralateral bias in face-responsivity under conditions of clutter in three-item displays (Fig. 5C), but we did replicate the well-known contralateral bias for isolated face stimuli (Fig. 2C; Silson et al., 2021; Groen et al., 2022). This observation serves as a reminder that anything we learn about the response profile of the face-selective network or the processing of signals in ITC under typical experimental conditions of fixating a single stimulus presented in isolation in the visual field needs to be tested for contextual tolerance (Wardle and Baker, 2020) and generalization to more naturalistic viewing conditions.

References


