Distinct yet proximal face- and body-selective brain regions enable clutter-tolerant representations of the face, body and whole person.

Libi Kliger[1] & Galit Yovel[1,2]

[1] The School of Psychological Sciences and [2] Sagol School of Neuroscience,

Tel Aviv University, P.O. Box 39040, Tel Aviv 6997801, Israel

Abbreviated title: ***Neural mechanism for person perception in clutter***

Correspondence:

Libi Kliger, libi.kliger@gmail.com

Galit Yovel gality@tauex.tau.ac.il

Number of pages: 34

Number of figures: 7

Number of words for abstract: 208, introduction: 632, and discussion: 1499.

Conflict of interest statement

The authors declare that there is no conflict of interest.

Acknowledgments

This work is supported by a grant from the Israeli Science Foundation (ISF 446/16).

Abstract

Faces and bodies are processed in separate but adjacent regions in primate visual cortex. Yet, the functional significance of dividing the whole person into areas dedicated to its face and body components and their neighboring locations remains unknown. Here we hypothesized that this separation and proximity together with a normalization mechanism generate clutter-tolerant representations of the face, body and whole person when presented in complex multi-category scenes. To test this hypothesis, we conducted a fMRI study, presenting images of a person within a multi-category scene to human male and female participants and assessed the contribution of each component to the response to the scene. Our results revealed a clutter-tolerant representation of the whole person in areas selective for both faces and bodies, typically located at the border between the two category-selective regions. Regions exclusively selective for faces or bodies demonstrated clutter-tolerant representations of their preferred category, corroborating earlier findings. Thus, the adjacent locations of face- and body-selective areas enable a hardwired machinery for decluttering of the whole person, without the need for a dedicated population of person-selective neurons. Thus, the distinct yet proximal functional organization of category-selective brain regions enhances the representation of the socially significant whole person, along with its face and body components, within multi-category scenes.

Significance statement

It is well established that faces and bodies are processed by dedicated brain areas that reside in nearby locations in primates' high-level visual cortex. However, the functional significance of the division of the whole person to its face and body components, their neighboring locations and the absence of a distinct neuronal population selective for the meaningful whole person remained puzzling. Here we proposed a unified solution to these fundamental open questions. We show that consistent with predictions of a normalization mechanism, this functional organization enables a hardwired machinery for decluttering the face, body and the whole person. This generates enhanced processing for the socially meaningful whole person and its significant face and body components in multi-category scenes.

Introduction

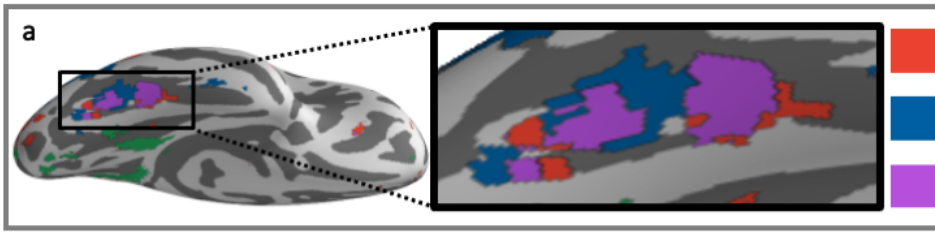
Intact processing of faces and bodies is critical for effective social interactions. The functional separation and the anatomical proximity of face and body-selective brain areas in human and monkey high-level visual cortex is well-established (Foster et al., 2021; Harry et al., 2016; Pinsk et al., 2005, 2009; Premereur et al., 2016; Schwarzlose et al., 2005; Weiner & Grill-Spector, 2013; Zafirova et al., 2022). Yet the functional significance of this anatomical organization remained unclear (for recent reviews see (Hu et al., 2020; Taubert et al., 2022). Why are faces and bodies processed by dedicated distinct mechanisms, despite their natural co-occurrence in the whole person? Why despite the significance of the whole-person for social perception, a distinct population of person-selective neurons/brain areas has not been reported so far? Why do face and body-selective regions reside in nearby locations? Here we propose a unified account for these questions. In particular, we test the hypothesis that this division of the whole person into distinct but proximally located face and body-selective areas supports the generation of clutter-tolerant representations for the face alone, the body alone or the whole person when presented in multi-category scenes (**Error! Reference source not found.**). An advantage of this mechanism is that it eliminates the need for an additional population of person-selective neurons.

A central challenge of the visual system is to generate a veridical representation of objects in multi-category scenes. This effect of clutter is reflected in a reduced neural response to an object when presented with other objects than when presented alone (Bao & Tsao, 2018; MacEvoy & Epstein, 2009, 2011; Miller et al., 1993; Rolls & Tovee, 1995; Zoccolan et al., 2005). Interestingly, this reduced response was not found in category-selective areas, where the response to the preferred category remained unaffected when presented with other categories (Bao & Tsao, 2018; Reddy & Kanwisher, 2007). A normalization mechanism was suggested to account for these

findings (Bao & Tsao, 2018; Carandini & Heeger, 2012; Heeger, 2011; Reynolds & Heeger, 2009). This mechanism posits that the response of a neuron is normalized by the response of its neighboring neurons (i.e., the normalization pool, **Error! Reference source not found.**b-c). Therefore, when a neuron is surrounded by neurons that are selective for its non-preferred categories, its response to simultaneous presentation of the two categories is reduced relative to the response to each object alone (Zoccolan et al., 2005). However, when the surrounding neurons are selective to the same category (i.e., a homogeneous normalization pool), as typically found in category-selective regions, the response to a preferred and a non-preferred stimuli presented together is similar to the response to the preferred stimulus presented alone (i.e., a max response). This essentially generates a clutter-tolerant representation of the preferred category (**Error! Reference source not found.**d-e) (Bao & Tsao, 2018; Kliger & Yovel, 2020, Reddy & Kanwisher, 2007). These findings offer a mechanistic account for the advantage of clustering neurons that are selective for significant categories, such as faces or bodies.

Here we propose that the proximity of clusters of face- and body-selective neurons together with the same normalization mechanism further enables a hardwired clutter-tolerant representation of the whole person (**Error! Reference source not found.**f). This is enabled by the presence of both face-selective neurons and body-selective neurons in the normalization pool (see derivations of the normalization equations in Appendix), as typically found in the border between face and body-selective areas (Kliger & Yovel, 2020). To test this prediction, we presented the whole person in a multi-category scene and assessed whether the representations of the multi-category scene is biased to the whole person in areas that are selective to both the face and the body (Figure 1f). These findings would indicate that the adjacent locations of face and body-selective clusters of neurons generate a clutter-tolerant representation for the face alone, the body alone or the whole person when presented in multi-category scenes.

a



Face-selective voxels

Body-selective voxels

Face & body-selective voxels

Neuron response prediction – The Normalization Model

b

$$R_j(F+B+C+P) = \gamma \frac{F_j + B_j + C_j + P_j}{\sigma + \sum_k F_k + \sum_k B_k + \sum_k C_k + \sum_k P_k}$$

Neuron j: stimulus drive

Neuron j: Response to multi-category scene (Face, Body, Chair and Place)

Normalization pool: sum of responses of surrounding neurons

c

The neuron response can be represented with a mathematically equivalent equation (see Appendix) as a weighted contribution of the the isolated categories, with weights determined by the selectivity of the normalization pool:

$$R_j(F+B+C+P) = \beta_{Face} \cdot R_j(F) + \beta_{Body} \cdot R_j(F) + \beta_{Chair} \cdot R_j(C) + \beta_{Room} \cdot R_j(P)$$

$$\beta_{Face} = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

Sum of responses of surrounding neurons to the face

Sum of responses of surrounding neurons to all categories

d

Face-selective voxels

$\Sigma_k F_k \gg \Sigma_k B_k, \Sigma_k C_k, \Sigma_k P_k$ ⇨ $\beta_{Face} > \beta_{Body}, \beta_{Chair}, \beta_{Room}$

The normalization pool contains mainly face-selective neurons.

The response to the multi-category scene is biased towards the face, decluttering non-face stimuli.

e

Body-selective voxels

$\Sigma_k B_k \gg \Sigma_k F_k, \Sigma_k C_k, \Sigma_k P_k$ ⇨ $\beta_{Body} > \beta_{Face}, \beta_{Chair}, \beta_{Room}$

The normalization pool contains mainly body-selective neurons.

The response to the multi-category scene is biased towards the body, decluttering non-body stimuli.

f

Face & body-selective voxels

$\Sigma_k F_k, \Sigma_k B_k \gg \Sigma_k C_k, \Sigma_k P_k$ ⇨ $\beta_{Face}, \beta_{Body} > \beta_{Chair}, \beta_{Room}$

$\Sigma_k F_k \approx \Sigma_k B_k$ ⇨ $\beta_{Face} \approx \beta_{Body}$

The normalization pool contains face-selective neurons and body-selective neurons.

The response to the multi-category scene is biased towards the face and body (person), decluttering non-person stimuli.

Figure 1: Predicted response of single neurons to a multi-category scene in face and body-selective cortex. (a) The functional organization of face- and body-selective areas in the ventral temporal cortex in a representative subject in MNI space. Colors indicate category-selective voxels: voxels selective only to faces (red), only to bodies (blue) or to both faces and bodies in the border between them (purple). (b) A multi-category scene composed of a face, a body, a chair and a room, and the normalization equation representing the response of a single neuron to that scene. (c) A mathematical equivalent of (b), which shows the predicted contribution (weight, β) of each category to the response to the multi-category scene. According to this, the contribution of each category is determined by the sum of responses of the surrounding neurons (i.e., the normalization pool) to that category relative to the sum of responses of the surrounding neurons to all categories. Equation is shown for β_{Face}. See Appendix for detailed mathematical derivations and for the complete formulas for each of the β's. (d) A pictorial illustration of the predicted representations of the multi-category scene in voxels that contain a homogenous population of face-selective neurons. The normalization equation predicts that the response of a neuron to the scene will be biased to the response of the face. (e) Same as (c) for body-selective voxels. (f) A pictorial illustration of the predicted response to the multi-category scene in voxels that contain adjacent populations of face-selective and body-selective neurons. The normalization equation predicts a biased response to the face and the body, essentially filtering-out non-person stimuli. Predictions are made for the neuronal response but can also be estimated with fMRI (see Appendix).

Materials and Methods

Participants:

Fifteen healthy volunteers (3 women, ages 21-31, 1 left-handed) with normal or corrected-to-normal vision participated in both experiments. Participants were paid $15/hr. All participants provided written informed consent to take part in the study, which was approved by the ethics committees of the Sheba Medical Center and Tel Aviv University and performed in accordance with relevant guidelines and regulations. The sample size for each experiment (N = 15) chosen for this study was similar to the sample size of other fMRI studies that examined the representation of multiple objects in high-level visual cortex (10-15 subjects per experiment) (Baeck et al., 2013; Baldassano et al., 2016; Kaiser et al., 2014; Kaiser & Peelen, 2018; Kliger & Yovel, 2020; MacEvoy & Epstein, 2009, 2011; Reddy et al., 2009; Song et al., 2013)

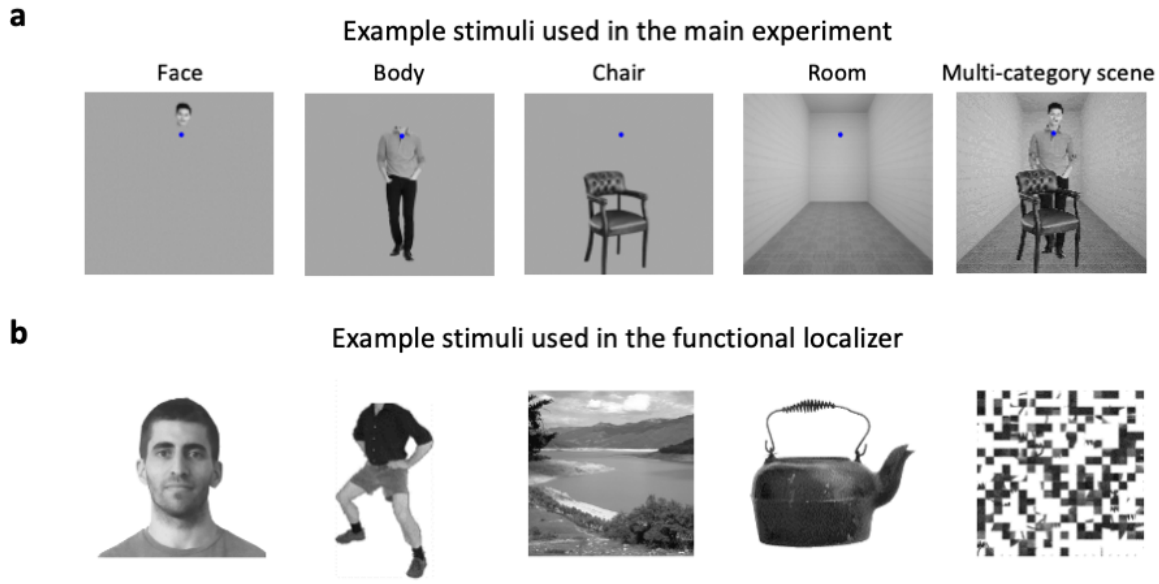


Figure 2: Example of the stimuli used in the experiment. (a) Stimuli of the main experiment, including isolated faces, bodies, chairs and rooms and multi-category scenes composed of all of the isolated categories. The isolated categories were placed in the same location in the visual field as they appeared in the multi-category scene, and subjects were instructed to maintain fixation on the blue fixation dot throughout the experiment. These stimuli were used to estimate the contribution of each of the isolated categories to the response to the multi-category scene. (b) Stimuli of the functional localizer. Stimuli included images of faces, bodies, outdoor scenes, non-living objects and scrambled objects. These stimuli were used to assess the magnitude of category selectivity of each voxel to each category.

Stimuli

Main experiment. The stimulus set consisted of grey-scale images of a multi-category scene as well as its isolated parts: Face, Body, Chair and Room (Figure 2a). The face and body stimuli were created by using 7 grey-scale images of a whole person standing in a straight frontal posture with the background removed, downloaded from the internet (Kliger & Yovel, 2020). Each image of a person was cut into two parts in the neck area resulting in a face stimulus and a headless-body stimulus for each identity. The chair stimuli included 7 images of chairs downloaded from the internet and scaled to a size that fits common proportions between a standing person and a chair. The face, body and chair stimuli were presented with a grey background. The room stimuli

included 7 empty rooms created using the website https://roomstyler.com/3dplanner and converted to greyscale. The contrast and luminance of the rooms were scaled such that they had the same contrast and luminance of a chair including its grey background. The complex scene stimuli included 7 images of a person inside a room, with the person standing behind a chair (preserving real-life proportion and composition between the person and the chair) at the center of the room. The single-category stimuli were presented at the exact same locations on the screen as they were presented within the multi-category scene. A fixation point was presented in the upper central part of all stimuli, at a location corresponding to the lower part of the neck of the standing person. The size of the multi-category scene images was 13.6X13.6 degrees of visual angle.

Functional localizer stimuli. Functional localizer stimuli were grey-scale images of faces, headless bodies, outdoor scenes, non-living objects and scrambled images of these objects (Figure 2b). Each category consisted of 80 different images. The size of the stimuli was approximately 5.5X5.5 degrees of visual angle.

Apparatus and Procedure

fMRI acquisition parameters. fMRI data were acquired in a 3T Siemens MAGNETOM Prisma MRI scanner in Tel Aviv University, using a 64-channel head coil. Echo-planar volumes were acquired with the following parameters: repetition time (TR) = 1 s, echo time = 34 ms, flip angle = 60°, 66 slices per TR, multi-band acceleration factor = 6, slice thickness = 2 mm, field of view = 20 cm and 100 × 100 matrix, resulting in a voxel size of 2 × 2 × 2 mm. Stimuli were presented with Matlab (The MathWorks Inc.) and Psychtoolbox (Brainard, 1997; Kleiner et al., 2007) and displayed on a 32" high-definition LCD screen (NordicNeuroLab) viewed by the participants at a distance of 155 cm through a mirror located in the scanner. Anatomical MPRAGE images were collected with 1 × 1 × 1 mm resolution, echo time = 2.45 ms, TR = 2.53 s.

Experimental procedure. The study included a single recording session with six runs of the main experiment and three runs of the functional localizer. Each of the six main-experiment runs included fifteen pseudo-randomized mini-blocks, three of each of the following experimental conditions: Face, Body, Chair, Room and the multi-category stimulus, as described in the stimuli section (see also Figure 2a). Each mini-block included eight stimuli, of which seven were of different images and one image repeated for the 1-back task. Each mini-block lasted 6 seconds and was followed by 12 seconds of fixation. A single stimulus display time was 0.375 seconds, and inter-stimulus-interval was 0.375 seconds. Subjects performed a 1-back task (one repeated stimulus in each block). Each run began with a six seconds (6 TRs) fixation (dummy scan) and lasted a total of 276 seconds (276 TRs). Subjects were instructed to maintain fixation throughout the run and their eye movements were recorded with an Eye tracker (EyeLink®).

Each functional localizer run included 21 blocks: 5 baseline fixation blocks and 4 blocks for each of the five experimental conditions: faces, bodies, nature-scenes, objects and scrambled objects. Each block presented 20 stimuli of 18 different images of which two were repeated for a 1-back task. Each stimulus was presented for 0.4 sec with 0.4-sec inter-stimulus interval. Each block lasted 16 seconds. Each run began with a 6-seconds fixation (6 TRs) and lasted a total of 342 seconds (342 TRs).

Data analyses

fMRI data analysis and preprocessing. fMRI analysis was performed using SPM12 software, Matlab (The MathWorks Inc.) and R (R Development Core Team, 2011) costumed scripts, Freesurfer (Dale et al., 1999), pysurfer (https://pysurfer.github.io) and Python (http://www.python.org) costumed scripts for the surface generation and presentation. The code that was used for data analyses is available at https://github.com/gylab-TAU/Multi_Category_Scenes_fMRI_analysis. The first six volumes in each run were acquired during a blank-screen display and were discarded from the analysis as “dummy scans”. The data

were then preprocessed using realignment to the mean of the functional volumes and co-registration to the anatomical image (rigid body transformation), followed by spatial normalization to MNI space. Spatial smoothing was performed for the localizer data only (5 mm). A GLM was performed with separate regressors for each run and each condition, including 24 nuisance-motion regressors for each run (6 rigid body motion transformation, 6 motion derivatives, 6 square of motion and 6 derivatives of the square of motion) and a baseline regressor for each run. In addition, a "scrubbing" method (Power et al., 2012) was applied for every volume with frame-displacement (FD) > 0.9 by adding a nuisance regressor with a value of 1 for that specific volume and zeros for all other volumes. Percent signal change (PSC) for each voxel was calculated for each experimental condition in each run by dividing the beta weight for that regressor by the beta weight of the baseline for that run.

Linear model fitting. Mean percent signal change (PSC) across runs to the face, the body, the room, the chair and the multi-category scene conditions from the main experiment data was extracted for each voxel of each subject. For each subject, we defined a moving mask of a sphere of 27 voxels (i.e., a 3x3x3 grid). We used a relatively small sphere of 27 voxels to assure that voxels within each sphere are homogenous in terms of their category selectivity. For each sphere, we fitted a linear model with its voxel data as features (i.e., the percent signal change, PSC, in each of these voxels) to predict the response to the multi-category scene based on the response to the isolated categories:

$$\text{MultiCategory } Scene = \beta_{Face} \cdot Face + \beta_{Body} \cdot Body + \beta_{Room} \cdot Room + \beta_{Chair} \cdot Chair + \varepsilon \quad (1)$$

The beta coefficients of these models represent the contribution of each of the isolated categories to the response to the multi-category scene of each sphere of each subject. Note that the beta coefficients of the multi-category response model are not the same as the betas derived from the standard fMRI GLM analysis. The betas from the latter analysis are used to determine the percent

signal change (PSC) to each of the single- and multi-category stimuli as a measure of the fMRI response to these stimuli.

Anatomical regions of interest (anatomical ROI) definition. We defined voxels that belong to the ventro-temporal cortex (VTC) and lateral-occipital cortex in the right hemisphere by using a mask based on the Harvard-Oxford Atlas (Desikan et al., 2006; Frazier et al., 2005; Goldstein et al., 2007; Makris et al., 2006). We used the max-probability mask (threshold=0) with voxel size of 2x2x2 mm. The ventro-temporal mask included the following areas from the Harvard-Oxford Atlas: Inferior Temporal Gyrus, posterior division, Inferior Temporal Gyrus, temporooccipital part, Parahippocampal Gyrus, anterior division, Parahippocampal Gyrus, posterior division, Temporal Fusiform Cortex, anterior division, Temporal Fusiform Cortex, posterior division, Temporal Occipital Fusiform Cortex, Occipital Fusiform Gyrus (labels 14-15, 33-34, 36-39, respectively). The lateral-occipital mask included the following areas from the Harvard-Oxford Atlas: Middle Temporal Gyrus, temporooccipital part, Lateral Occipital Cortex, superior division, Lateral Occipital Cortex, inferior division (labels 12, 21-22, respectively).

We selected the area labeled Frontal Pole (label 0) as a control non-visual area. The number of spheres that was randomly selected to be included in this control area was the average number of spheres of the category-selective areas for each participant.

Voxels definition by category-selectivity. Based on the functional localizer data, we estimated the selectivity of each voxel of individual subjects for the face and the body by using contrast t-maps of Face > Object, Body > object and Outdoor-scenes > Object, respectively. We used only these three categories since their definition shares the same baseline (i.e., they are all compared to Object). We excluded the general object-selective region since the common definition of these areas (Objects > Scrambled objects) will result in areas that are not category-specific but are similarly responsive to all object categories. Within each anatomical ROI (i.e., ventro-temporal cortex and lateral-occipital cortex) we defined several types of voxels based on their selectivity

for these three categories. Face-selective voxels were defined as voxels that are selective only for faces over objects ($p < .0001$) and to faces over bodies ($p < .0001$) and not selective for bodies and places ($p > .01$). These criteria assured that the majority of the neurons within these voxels are face-selective. One subject did not have face selective voxels (i.e., selective only for faces and not for other categories). Similarly, we defined body-selective voxels as voxels selective only for bodies over objects ($p < .0001$) and to bodies over faces ($p < .0001$) but not selective for faces or places over objects ($p > .01$). In addition, we defined face and body-selective voxels, by selecting voxels that are selective for both faces and bodies (but not for places). These voxels contain clusters of face-selective neurons and body-selective neurons. All participants showed voxels that were selective to both the face and the body In the ventral-temporal cortex. In the lateral-occipital cortex only 5 out of 15 participants showed voxels that are selective to both the face and the body. Because the main novel contribution of this work is the response of the ROI selective to both faces and bodies, we focused only on the ventral-temporal ROIs and did not include lateral-occipital ROIs in our analyses.

The contribution of each category to the multi-category scene representation. For each subject, we calculated the betas of the model from Equation 1 for spheres of 27 voxels in the category-selective areas described above (see model fitting description above). To reduce statistical dependency as a result of the overlapping moving mask, we calculated the mean using an interleaved mask, taking only spheres that their centers are not immediately adjacent to one another. We computed the mean beta across all spheres in an ROI for each participant and across participants. We calculated the variance inflation factor (VIF), which provides a measure of multicollinearity of the beta coefficients, and removed spheres in which the VIF was larger than 10. We then performed repeated measure ANOVAs to examine the contribution of the isolated categories to the multi-category scene for voxels selective for pairs of categories within the ventro-temporal cortex. We used Category (Face, Body, Room and Chair) and ROI selectivity (face-selective, body-selective and face and body selective) as within-subject factors and the beta

coefficients of the multi-category response model as a dependent variable. To test our specific hypothesis with respect to the representation of faces and bodies relative to the other categories in the different ROIs, we used paired t-test. One subject did not have face-selective voxels (i.e., selective only for faces and not for other categories) and one subject did not have body-selective voxels (i.e., selective only for bodies and not for other categories) based on the above-mentioned criteria and therefor was not included in statistical analysis that compared between these ROIs. The subjects were included in analysis of other category-selective areas.

Defining non-saturated voxels. To test whether the BOLD response to the multi-category scene was saturated, we conducted the following analysis. For each voxel, we compared the maximum response (PSC) to a single category and to the response to the multi-category scene:

$$\text{Non-saturated voxels: } \max\{Face_{PSC}, Body_{PSC}, Chair_{PSC}, Room_{PSC}\} > MultiCategory\ Scene_{PSC}$$

A non-saturated response to the multi-category scene is evident in voxels that show higher response to a single category than to a multi-category stimulus.

Predictions

We measured the fMRI response to a multi-category scene of a person in a room with a chair and to each of its isolated categories (see Figure 2 and Methods). The predictions of the response to the multi-category scene according to the normalization model in the different category-selective voxels are specified in Figure 1 (see Appendix for complete mathematical derivations and predictions). The normalization model predicts that voxels that are selective for either the face or the body - therefore containing one homogenous population of face or body-selective neurons - the representation of multi-category scenes will be biased to the preferred category, decluttering non-preferred stimuli in the scene. In addition, in voxels that are selective to both the face and the body - therefore containing two homogenous populations of face and body-selective neurons, the

representation of multi-category scenes will be biased to both preferred face and body categories, decluttering non-person stimuli (chair and room) within the scene.

Results

A clutter-tolerant representation of the whole person in face- and body-selective areas

We defined three types of voxels in the ventro-temporal area based on their selectivity for the isolated categories (see methods): (1) face voxels: voxels selective for faces but not for bodies, non-living objects or places; (2) body voxels: voxels selective for bodies but not for faces, non-living objects or places; and (3) face-body selective voxels: voxels selective for both faces and bodies but not for non-living objects and places (usually located at the border between face and body areas) (see Methods and Figure 1). To assess the contribution of each category to the representation of the multi-category scene, subjects viewed a different, independent set of stimuli containing a multi-category visual scene of a whole person standing next to a chair located inside a room, as well as stimuli of each of the components of the scene shown separately (Figure 2a). We extracted the mean percent signal change (PSC) response from each voxel for each of the isolated-component stimuli. We then used these voxel-level PSCs of each component of the multi-scene as predictors for a linear model, and the PSC response for the multi-category scene as the predicted variable:

$$\mathrm{Scene}_{\mathrm{PSC}} = \beta_{Face} \cdot Face_{PSC} + \beta_{Body} \cdot Body_{PSC} + \beta_{Room} \cdot Room_{PSC} + \beta_{Chair} \cdot Chair_{PSC} + \varepsilon \quad (1)$$

The beta coefficients of the above model represent the contribution of each of the isolated categories to the response to the multi-category scene. For each subject, we defined a moving mask of a sphere of 27 (3x3x3) voxels. For each sphere, we fitted a linear model with its voxel data as features to predict the response to multi-category scene from the response to each its component. We included only interleaved spheres to avoid high overlap and statistical

dependency between overlapping spheres. Note that the beta coefficients of the multi-category response model indicate the predicted contribution of each category to the fMRI response to the multi-category scene, not the betas derived from the standard fMRI GLM analysis.

Figure 3a-c depicts the contribution of each of the isolated categories to the response of the multi-category scene (i.e., the beta coefficients of the above linear model in (1) for the face-selective, body-selective and face-body selective voxels of each participant and averaged across participants. We performed a repeated measure ANOVA with category (face, body, chair and room) and voxel selectivity (face-selective voxels, body-selective voxels and face and body selective voxels) as within-subject factors and the contribution to the complex scene representation (i.e., the beta coefficients of the linear model) as the dependent variable. We found a significant effect for Category [$F(3,36) = 35.363$, $p < 0.0001$, $\eta_G^2 = 0.526$] as well as a significant category x voxel selectivity interaction [$F(6,72) = 8.625$, $p < 0.0001$, $\eta_G^2 = 0.275$].

We performed paired t-tests to compare the contribution of the preferred relative to the non-preferred categories. In line with the predictions of the normalization model (Figure 1), the contribution of the face to the representation of the multi-category scene in face-selective voxels was higher than the contribution of each of the other categories [$\beta_{Face} - \beta_{Body}$: mean = 0.264, $t(12) = 3.274$, $p = 0.0200$, Cohen's d = 0.908, 95% C.I. (0.088, 0.440); $\beta_{Face} - \beta_{Room}$: mean = 0.452, $t(12) = 6.707$, $p < 0.0001$, Cohen's d = 1.860, 95% C.I. (0.305, 0.598); $\beta_{Face} - \beta_{Chair}$: mean = 0.588, $t(12) = 6.347$, $p = 0.0001$, Cohen's d = 1.760, 95% C.I. (0.386, 0.790); all p values are corrected for multiple comparisons] (Figure 3a). Similarly the contribution of the body was higher than the contribution of each of the other categories in the body-selective voxels categories [$\beta_{Body} - \beta_{Face}$: mean = 0.366, $t(12) = 5.443$, $p = 0.0004$, Cohen's d = 1.510, 95% C.I. (0.220, 0.513); $\beta_{Body} - \beta_{Room}$: mean = 0.422, $t(12) = 5.226$, $p = 0.0006$, Cohen's d = 1.450, 95% C.I.

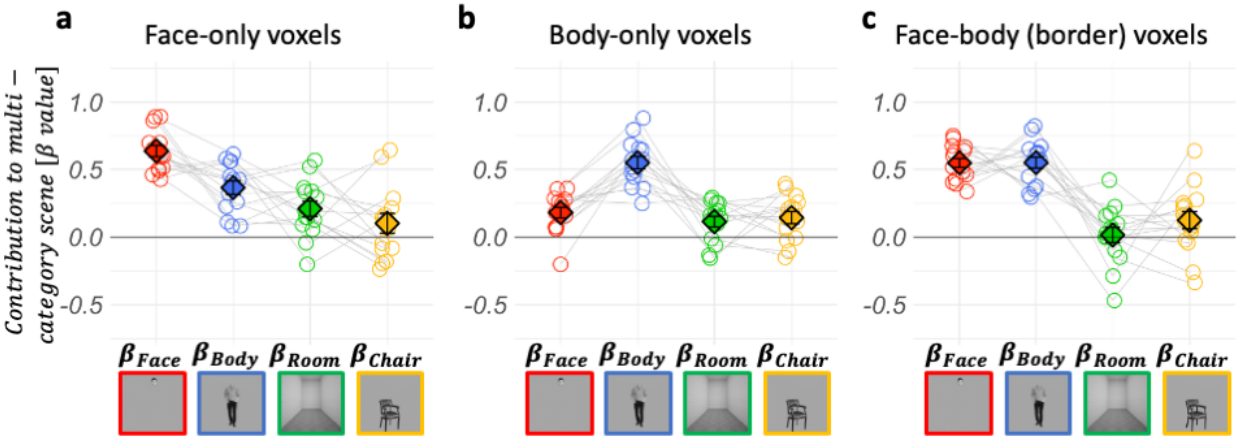


Figure 3: The contribution of single categories to the representation of a multi-category scene in face- and body-selective voxels. (a-c) The contribution of each isolated category to the representation of the multi-category scene in the right ventro-temporal face- and body-selective voxels as depicted by the β coefficients of the linear model (Equation 1) in (a) face-selective voxels, (b) body-selective voxels, (c) face- and body-selective voxels. Each dot indicates the mean β of a single subject. Gray lines connect the β's of the same subject. Diamonds and error bars indicate the group mean and s.e.m., respectively. Note that the β's of the linear model are not the betas extracted from the GLM that evaluates the correspondence between the fMRI hemodynamic response and stimulus presentation but indicate the contribution of each category to the response to the multi-category scene.

(0.246, 0.598); $\beta_{Body} - \beta_{Chair}$: mean = 0.421, t(12) = 4.979, p = 0.0010, Cohen's d = 1.381, 95% C.I. (0.237, 0.606); all p values are corrected for multiple comparisons] (Figure 3b). Finally, the contribution of the preferred face and body was higher than the contribution of each of the non-preferred categories in the face and body selective voxels [$(\beta_{Face} + \beta_{Body})/2 - \beta_{\mathrm{Room}}$: mean = 0.516, t(12) = 8.729, p < 0.0001, Cohen's d = 2.421, 95% C.I. (0.387, 0.645); $(\beta_{Face} + \beta_{Body})/2 - \beta_{\mathrm{Chair}}$: mean = 0.447, t(12) = 5.789, p = 0.0003, Cohen's d = 1.605, 95% C.I. (0.279, 0.615) ; all p values are corrected for multiple comparisons] (Figure 3c). The difference between the contribution of the face and the body in these voxels was not statistically significant [$\beta_{Face} - \beta_{Body}$: mean = -0.012, t(12) = -0.164, p = 0.872 (not corrected), Cohen's d = -0.046, 95% C.I. (-0.1707, 0.146)]. When comparing the null model against the alternative using Bayesian t-test, the null model is preferred over the alternative [BF =0.282]. In addition, the sum of betas within each type

of voxels is a little over 1, as predicted by the mathematical model (see Appendix 1) [$\beta_{Face} + \beta_{Body} + \beta_{room} + \beta_{Chair}$: Face only: mean = 1.302, sd = 0.301; Body only: mean = 1.012, sd = 0.128; Face-body: mean = 1.268; sd = 0.222].

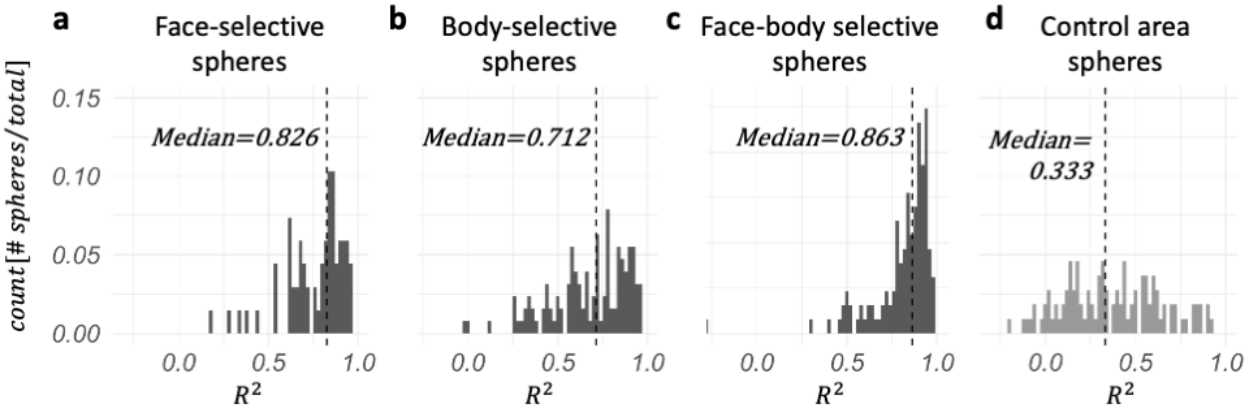


Figure 4: The distribution and median of R^2 values of all linear models (Equation 1) calculated for each sphere and subject for (a) face-selective voxels, (b) body-selective voxels, (c) face- and body-selective voxels (d) control non-visual area in the frontal pole.

To assess goodness of fit of the normalization model to the response to the multi-category scene, we computed the R^2 for each sphere. Figure 4a-c shows the distribution of the R^2 values in face- and body-selective areas. The overall median R^2 = 0.817 indicates a good fit of the proposed model to the data from these areas. For comparison, we defined a control area in the frontal lobe that does not show visual category-selectivity (the frontal pole, see Methods). The R^2 in this region was much lower median R^2 = 0.333 (Figure 4d), indicating that, consistent with our predictions, the normalization model accounts for the response in visual category-selective cortex.

Results reported so far show that the response of voxels that are selective to both faces and bodies show a different pattern of response than voxels that are either face or body selective. To further demonstrate that the three ROIs show distinct response characteristics, we plotted a scatterplot of the difference between β_{Face} and β_{Body} (i.e., difference in contribution of the face and the body to the response to the multi-category scene) over the relative selectivity for faces

and for bodies, which was measured independently by the functional localizer, for all spheres of all subjects (Figure 5). This scatterplot demonstrates that the bias towards either face or body in the response to the multi-category scene is associated with the selectivity to these categories. Moreover, it shows the equal contribution of the face and the body to the response to the multi-category scene is a feature of a sub-population of voxels that are selective to both the face and the body.

We performed a mixed-model linear regression with a random intercept for subjects and found a positive association between $\beta_{Face} - \beta_{Body}$ and the relative selectivity of Face>Body [χ^2=68.372 when comparing to the null model, p<0.0001; Fixed effect estimates: slope=0.041, s.d.=0.005, t=8.697, p<0.0001; intercept=-0.059, s.d.=0.040, t=-1.481, p=0.158]. Similar finding were reported by Kliger & Yovel (2020) who presented the face, body and whole person in isolation

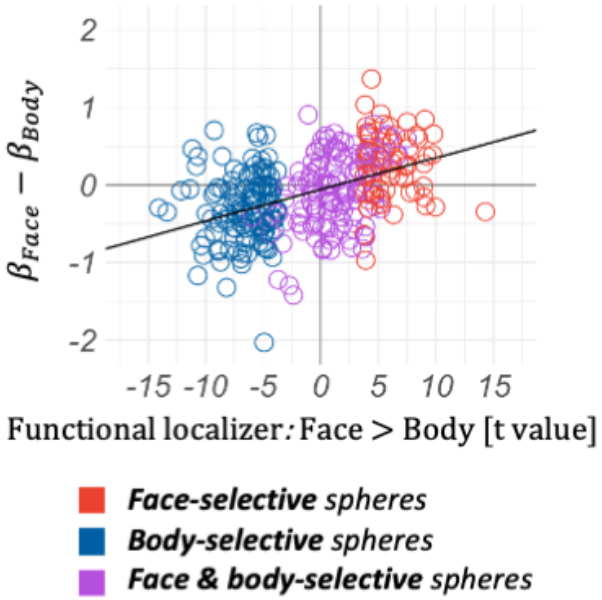


Figure 5: The contribution of face and body to multi-category scene co-varies with face-body selectivity. The scatterplot shows the co-variability of the difference in the contribution of the face and the body to the multi-category scene (y-axis), and the difference between face and body-selectivity that was independently measured by the functional localizer (x-axis) for each sphere of all subjects.

Taken together, these findings are consistent with predictions of the normalization model (**Error! Reference source not found.**), indicating that voxels that are selective for both the face and the body generate a representation that is biased to the whole person, while voxels that are selective for either of the single categories generate a representation that is biased to their preferred category.

Testing an alternative account to the normalization model: A summation model

An alternative summation account for our findings, which does not rely on the assumption of normalization, suggests that if a mixture of category-selective neurons responds independently, their combined response would be the sum of their individual responses to their preferred categories. This would lead to each beta values in the model (Equation 1) to be equal to 1. Our results do not support this summation prediction (Figure 3). Still, a lower than the sum response to the multi-category scene may be due to saturation of the BOLD signal, rather than lack of support for a summation model. However, under the summation model, the response to the single categories would never be higher than the response to the multi-category scene. We therefore examined whether such a pattern of activation exists. A voxel-wise analysis reveals a large proportion of voxels that show this pattern of response. We found that, in contrast to the summation account, 65.74% of the voxels in VTC that are selective to faces or bodies [face-selective: 61.76%; body selective: 77.95%; face-body selective: 55.81%] showed a higher response to a single-category relative to the response to the multi-category scene (Figure 6a-c). Subjects who did not have enough voxels in specific areas were excluded from Anova that examined all ROIs, but whenever available, their data are shown in Figure 6d-i and tested for planned t-test (two additional subjects with no face-only voxels, one additional subject with no body-only voxels, three subjects with no face-body voxels).

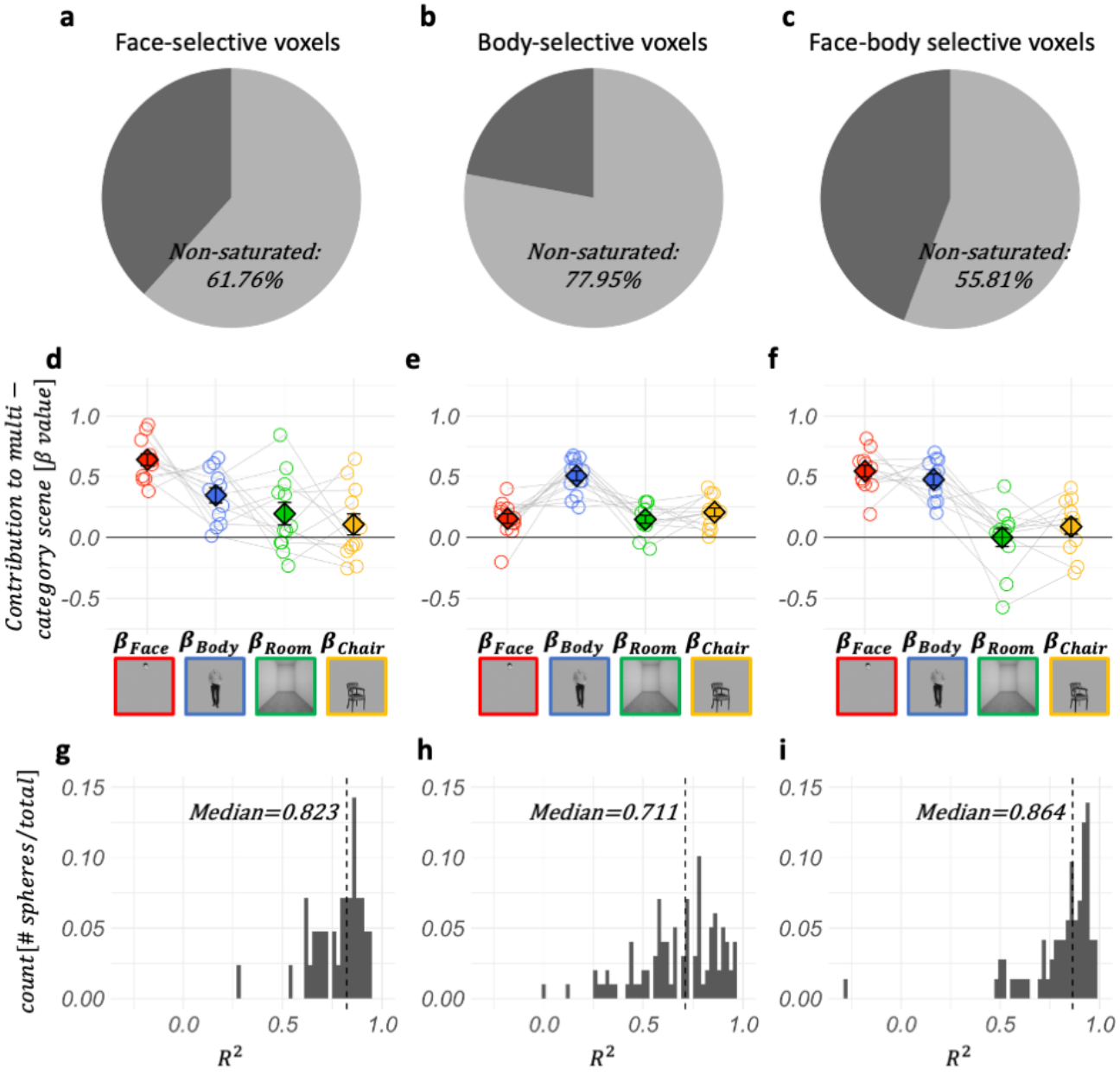


Figure 6: The contribution of single categories to the representation of a multi-category scene in non-saturated face- and body-selective voxels. (a-c) The proportion of voxels that show a higher response to single category than multi-category scenes (non-saturated voxels) in (a) face-selective, (b) body-selective and (c) face and body-selective areas. (d-f) The contribution of each isolated category to the representation of the multi-category scene in face- and body-selective voxels as depicted by the β's of the linear model (equation 1) in (a) non-saturated face-selective voxels, (b) non-saturated body-selective voxels, (c) non-saturated face and body selective voxels. Each dot indicates the mean β of a single subject. Gray lines connect the β's of the same subject. Diamonds and error bars indicate the group mean and s.e.m., respectively. (g-i) The distribution and median of R^2 values of all models (equation 1) for all spheres and all subjects for (g) face-selective voxels, (h) face-selective voxels, (i) face and body-selective voxels

We performed a similar analysis to the one described above (Figure 3), only for the non-saturated voxels – voxels that showed higher response to a single category than multi-category stimuli. Figure 6d-f depicts the contribution of each of the isolated categories to the response of the multi-category scene (i.e., the beta coefficients of the above linear model in (1) for only non-saturated voxels that are face-selective, body-selective and face-body selective of each participant and averaged across participants. Results are similar to the previous analysis (Figure 3). Figure 6g-i depicts the distribution of R^2 values of all models calculated for all spheres centered in non-saturated voxels for all subjects. The R^2 values (median = 0.790) were as high as those reported above for all voxels. Statistical analysis yielded similar results as well. A repeated measure ANOVA showed a significant effect for category [F(3,21) = 13.176, p < 0.0001, η_G^2 = 0.422] as well as a significant category x voxel selectivity interaction [F(6,42) = 5.699, p <=0.0002, η_G^2 = 0.277]. Additionally, the contribution of preferred categories was also significantly higher than the contribution of non-preferred categories. We performed paired t-tests to compare the contribution of the preferred relative to the non-preferred categories. In line with the predictions of the normalization model (Figure 1), the contribution of the face to the representation of the multi-category scene was higher than the contribution of all other categories in face-selective voxels, [$\beta_{Face} - \beta_{Body}$: mean = 0.293, t(11) = 2.83, p = 0.0491, Cohen's d = 0.817, 95% C.I. (0.065, 0.522); $\beta_{Face} - \beta_{Room}$: mean = 0.445, t(11) = 4.337, p = 0.0035, Cohen's d = 1.252, 95% C.I. (0.219, 0.670); $\beta_{Face} - \beta_{Chair}$: mean = 0.533, t(11) = 4.70, p = 0.0019, Cohen's d = 1.357, 95% C.I. (0.283, 0.782); all p values are corrected for multiple comparisons] (Figure 6d). Similarly, the contribution of the body was higher than the contribution of all other categories in the body-selective voxels [$\beta_{Body} - \beta_{Face}$: mean = 0.352, t(12) = 4.875, p = 0.0011, Cohen's d = 1.351, 95% C.I. (0.194, 0.509); $\beta_{Body} - \beta_{Room}$: mean = 0.357, t(12) = 5.363, p = 0.0005, Cohen's d = 1.487, 95% C.I. (0.212, 0.503); [$\beta_{Body} - \beta_{Chair}$: mean = 0.299, t(12) = 5.593, p = 0.0004, Cohen's d = 1.551, 95% C.I. (0.182, 0.415); all p values are corrected for multiple comparisons] (Figure 6e).

For the face and body selective voxels, the contribution of the preferred face and body was higher than the contribution of the non-preferred categories [$(\beta_{Face} + \beta_{Body})/2 - \beta_{\mathrm{Room}}$: mean = 0.511, t(11) = 6.081, p = 0.0002, Cohen's d = 1.755, 95% C.I. (0.326, 0.696); $(\beta_{Face} + \beta_{Body})/2 - \beta_{\mathrm{Chair}}$: mean = 0.422, t(11) = 5.485, p = 0.0006, Cohen's d = 1.583, 95% C.I. (0.253, 0.591) ; all p values are corrected for multiple comparisons] (Figure 6f). The difference between the contribution of the face and the body in these voxels was not statistically significant [$\beta_{Face} - \beta_{Body}$: mean = 0.069, t(11) = 0.783, p = 0.450 (not corrected), Cohen's d = 0.226, 95% C.I. (-0.124, 0.261)]. When comparing the null model against the alternative using Bayesian t-test, the null model is preferred over the alternative [BF =0.373]. These findings are consistent with prediction of the normalization model (**Error! Reference source not found.**) and results reported above that include all data (Figure 3), indicating that a simple summation account does not support the observed findings.

Finally, we examined behavioral measures during the fMRI scanning for the different stimulus categories. We measured performance on the 1-back task across the different categories. Performance was at ceiling for all categories. The mean accuracy was: Multi-category scene=0.99 (sd=0.010); Face = 0.98 (sd=0.018); Body=0.98 (sd=0.014); Room 0.97 (sd=0.019); Chair=0.98 (sd=0.015). In addition, we displayed the eye fixation patterns for each category. Figure 7 shows an overall similar pattern of fixations across the different stimuli, indicating that participants followed the instructions to focus on the fixation dot that was presented in the same location on screen across the different conditions during the experiment.

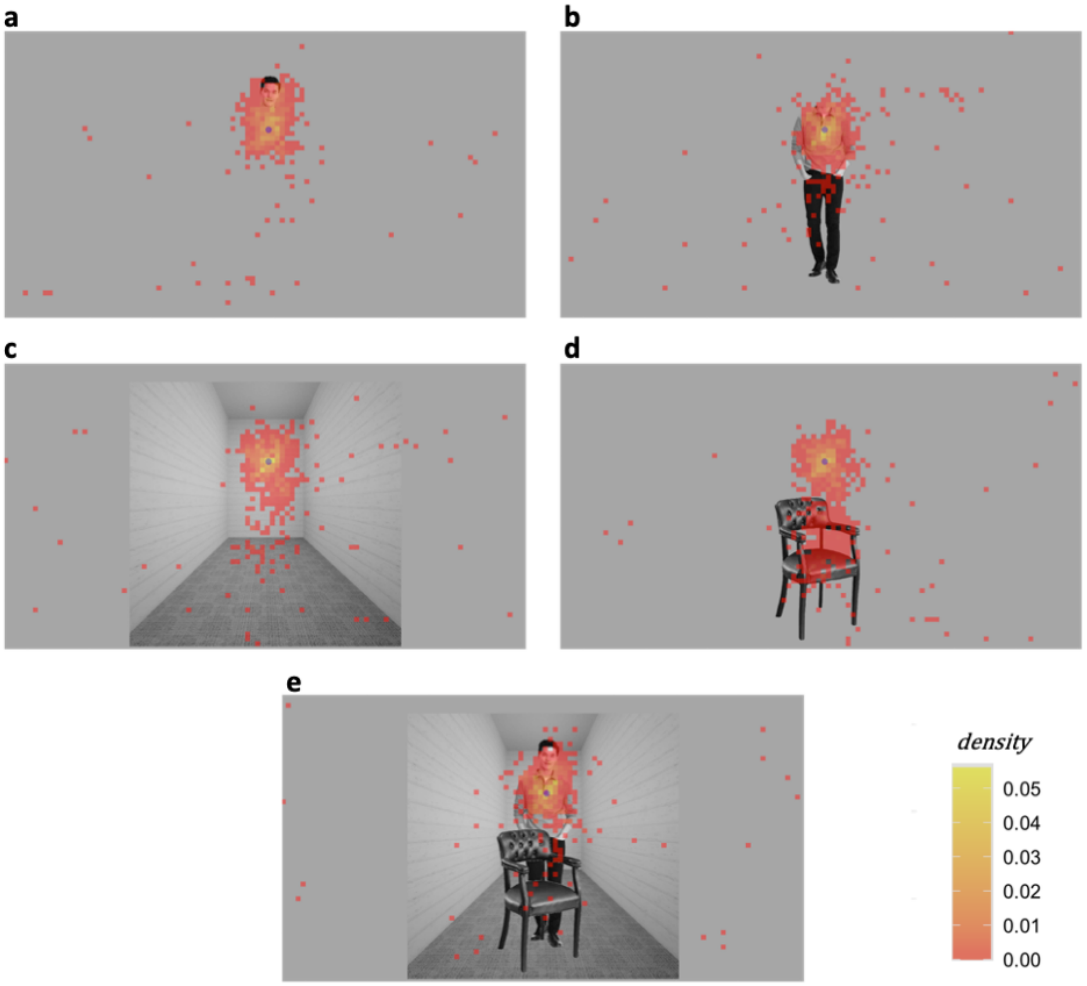


Figure 7: Eye tracker fixation patterns. Hit map of fixations duration along all experiment trials of all subjects. Background image is the picture with original location, size and fixation point as was presented in the experiment. The data is displayed for the different experimental conditions: (a) Face, (b) Body, (c) Room, (d) Chair, (e) Multi-category scene.

Discussion

The functional properties of face and body-selective areas have been extensively investigated in numerous neuroimaging and neurophysiological studies in humans and monkeys in the past two and a half decades. Still, the functional significance of their separation and adjacent locations remained unclear (for recent reviews see, Hu et al., 2020; Taubert et al., 2022). Our study

provides a mechanistic account for this long-standing puzzle by considering the operation of a well-established normalization model on distinct face and body-selective regions that reside in proximal anatomical locations. Consistent with predictions of the normalization model (see **Error! Reference source not found.** and mathematical derivations in the Appendix), we found that the representation of the multi-category scene was dominated by the face in face-selective areas, by the body in body-selective areas, and by both the face and body (i.e., the whole person) in the border between the face-selective and body-selective areas that is selective to both categories, filtering out non-preferred categories. To reveal this pattern of response our study presented the whole person within a multi-category scene (see Figure 1), unlike previous studies that presented an isolated face, body or whole person (Bernstein et al., 2014; Fisher & Freiwald, 2015; Kaiser et al., 2014; Kliger & Yovel, 2020; Song et al., 2013; Zafirova et al., 2022). This enabled us to measure the relative contribution of both preferred face and body categories and non-preferred categories to the multi-category scene.

Consistent with predications of the normalization model (Figure 1), we found that the proximal location of face and body-selective areas enables the generation of a clutter-tolerant representation of the meaningful combination of the face and body (the whole person). This machinery eliminates the need for a dedicated population of neurons that are selective to the combined whole-person stimulus. The generation of a clutter-tolerant representation of the whole person in neighboring face and body areas is accomplished through the same normalization mechanism that declutters preferred single categories within their category-selective cortex (see for example: Bao & Tsao, 2018; Kliger & Yovel, 2020; Reddy et al., 2009). According to the normalization model, when the normalization pool of a face-selective neuron is selective to bodies (or when the normalization pool of a body-selective neuron is selective to faces), the response to the multi-category scene will be a weighted mean of the response to the two categories, and reduced response to the non-preferred categories (see Appendix for mathematical derivations),

essentially generating a clutter-tolerant representation of the whole person. Furthermore, we demonstrated that an alternative model, which predicts summation instead of normalization, was not supported by the data.

Voxels that are selective to both the face and body typically reside in the border between the face and body-selective areas. Whereas our design does not allow us to determine whether these voxels contain two populations of face and body-selective neurons or one population of person-selective neurons, previous studies suggest that the former might be the case. Kaiser and colleagues (2014) used multi-voxel pattern analysis to ask if a person-selective region (a region that shows higher response to person than object stimuli) is composed of one population of person-selective neurons or nearby face-selective and body-selective neurons. Their results support the latter conclusion, though as noted by the authors, they do not provide conclusive evidence for the absence of person-selective neurons in this region. It should be noted that neurons responsive to the whole person were reported in the upper bank of the STS (Wachsmuth et al., 1994, see also fMRI findings by Fisher & Freiwald, 2015). This upper bank of the monkey STS may be the homologue of human STS (Yovel & Freiwald, 2013), which is known to show selectivity to biological motion of the whole person (Thompson et al., 2005). Another recent study found neurons responding to the face and body in a patch that is selective to whole person natural configuration in anterior IT (Zafirova et al., 2024). Our findings specifically focus on the border between the face and body-selective areas in the posterior ventral temporal cortex, where to the best of our knowledge such person-selective neurons were not reported. Still, future studies that will record neurons that reside in the border between face- and body-selective regions are required to further support this claim.

In a recent extensive review of the literature on the response of visual cortex to faces, bodies and the whole person, Taubert and colleagues (2022) have proposed four hypotheses regarding the organization of face and body-selective regions including separate networks, weakly-integrated

networks, strongly integrated networks or a single network. They concluded that current data do not fully support any of these hypotheses and called for future studies that will combine the face and body to address these open questions. Our study goes beyond this suggestion by predicting that the significance of this functional organization will be evident when the whole person is presented in multi-category scenes. Thus, our findings show the benefit of processing faces and bodies by both separated networks for enhancing the representations of the face or the body as well as integrated networks for enhancing the representation of the whole person.

The question of whether faces and bodies are processed by separated or integrated systems was also discussed in a recent review by Hu and Colleagues (Hu et al., 2020). According to their suggested model, faces and bodies are processed separately in posterior brain areas (OFA and EBA) but are integrated to the whole person in more anterior regions, the FFA and FBA (Bernstein et al., 2014; Fisher & Freiwald, 2015; Song et al., 2013). This model is consistent with our findings as the face and body adjacent voxels are primarily located in the fusiform gyrus (Figure 1), whereas more lateral and posterior face- and body-selective voxels (i.e., OFA and EBA) are located more distant from each other. Indeed, we found that the face and body selective regions were proximal only in third of the participants in the lateral-occipital cortex, whereas in ventral temporal cortex face and body regions reside adjacently in all our participants (see also Weiner & Grill-Spector, 2010, 2013). We suggest that functional organization enables such independent and integrated processing of the face and body, either by clusters of neurons that are located more remotely from one another (mostly posteriorly) or by nearby face and body regions (mostly ventrally), respectively.

The well-established organization of category-selective visual cortex has generated different hypotheses with respect to their emergence in particular locations in the high-level visual cortex (Deen et al., 2017; Saygin et al., 2016; van den Hurk et al., 2017). Recently, Op de Beeck and colleagues (2019) proposed three main factors that determine where category-selective areas

emerge in the visual cortex: (1) pre-existing feature selectivity; (2) computational hierarchy; and (3) domain-specific connectivity to areas outside the visual stream. The current study goes beyond the representation of single categories and highlights the benefit of positioning different category-selective regions in proximity, in particular to the representation of multi-category scenes. Theories that attempt to account for the pre-determined locations of category-selective areas should also consider the functional significance of their relative proximity for resolving the computational challenges of representing multi-category scenes.

The present study proposes a bottom-up mechanism that can bias the response to certain, significant categories by clustering homogenous category-selective neurons. Yet, other mechanisms have been suggested to bias the response to specific stimuli. For example, bottom-up, stimulus-driven mechanisms based on stimulus saliency can allocate resources toward a specific target (Beck & Kastner, 2005). Furthermore, a normalization operation was also shown to account for top-down mechanisms of selective attention that resolves competition among multiple stimuli (Desimone & Duncan, 1995; Reddy et al., 2009; Reynolds & Heeger, 2009) Thus, the proposed hardwired mechanism acts in concordance with other bottom-up and top-down mechanisms to resolve the challenge of processing rich, multi-category scenes (McMains & Kastner, 2011; Pessoa et al., 2003).

The biased representation to the whole person that we revealed is in line with behavioral studies that reported evidence for preferred processing of the whole person. For example, Mayer and colleagues (2015) showed that stimuli of the whole person pop out in cluttered scenes relative to other non-human stimuli. Downing et al. (2004) showed that they capture attention even when they are unattended; Privileged detection for whole person and faces relative to objects was also found in continuous flash suppression tasks (Stein et al., 2012). The clutter-tolerant representation that we revealed here for the whole-person may underlie these behavioral effects.

To summarize, our study offers a unified mechanistic account for long-standing questions about the neural representations of the face, the body and the whole person in high-level visual cortex. We explain how the same normalization mechanism enables the generation of a clutter-tolerant representation of each socially significant component (face or body) and their meaningful combination (whole person), thanks to the neighboring cortical locations of distinct clusters of face and body-selective neurons. More generally, our study reveals a new mechanism that is used by the visual system to resolve the challenging task of processing socially meaningful stimuli in cluttered scenes.

Appendix: Mathematical derivations of a model predicting the representation of a multi-category scene composed of four categories

According to the normalization model (Reynolds & Heeger, 2009), the measured neuronal response of a specific neuron (i.e., neuron j) to multi-category stimuli is divided by the sum of the responses of the surrounding neurons. This can be described by the following equation for a multi-category stimulus composed of four categories, denoted by A-D, such as the stimulus shown in Figure 1b, given by the following equation:

$$R_j(A+B+C+D) = \gamma \frac{A_j + B_j + C_j + D_j}{\sigma + \Sigma_k A_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k D_k},$$

where the measured response of a specific neuron (i.e., neuron j) to the stimuli presented together, $R_j(A+B+C+D)$, equals the response of the neuron to the sum of the stimuli, $A_j + B_j + C_j + D_j$, divided by the sum of the responses of the surrounding neurons (the normalization pool) to the stimuli, $\Sigma_k A_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k D_k$, and a constant, σ. The constants γ and σ are free parameters that are fitted to the data

We apply this equation to the multi-category scene that we used in our study, which is composed of a person, a chair and a room. Therefore, we have four categories presented in this multi-category scene, denoted as follows: a face (F), a body (B), a chair (C) and a room (i.e., a place, P):

$$R_j(F+B+C+P) = \gamma \frac{F_j + B_j + C_j + P_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}.$$

We can rewrite the normalization equation to express the response to a multi-category scene as a linear combination of the responses to each of the isolated categories composing the scene (Figure 1c).

We can separate the right side of the equation into two parts, yielding

$$R_j(F+B+C+P) = \frac{\gamma F_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k} + \frac{\gamma(B_j + C_j + P_j)}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}.$$

Next, we multiply the face part by a term equal to 1:

$$R_j(F+B+C+P) =$$

$$= \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k} \cdot \frac{\gamma F_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k} + \frac{\gamma(B_j + C_j + P_j)}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}.$$

Rewriting the equation, it becomes:

$$R_j(F+B+C+P) =$$

$$= \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k} \cdot \frac{\gamma F_j}{\sigma + \Sigma_k F_k} + \frac{\gamma(B_j + C_j + P_j)}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}.$$

Since the response to the isolated face according to the normalization model is given by:

$$R_j(F) = \frac{\gamma F_j}{\sigma + \Sigma_k F_k}$$

the response to the multi-category scene becomes:

$$R_j(F+B+C+P) = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k} \cdot R_j(F) + \frac{\gamma(B_j + C_j + P_j)}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}.$$

We can write this equation as

$$R_j(F+B+C+P) = \beta_{Face} \cdot R_j(F) + \frac{\gamma(B_j + C_j + P_j)}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k},$$

where

$$\beta_{Face} = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

is the coefficient of the face in this linear combination. Note that the coefficient depends only on the selectivity of the surrounding neurons (i.e., the normalization pool) for the multi-category scene that was presented, and not on the selectivity of neuron j itself.

For simplicity we showed only the derivations for the face weight, but similar derivations can be performed for all other categories, yielding:

$$R_j(F+B+C+P) = \beta_{Face} \cdot R_j(F) + \beta_{Body} \cdot R_j(F) + \beta_{Chair} \cdot R_j(C) + \beta_{Place} \cdot R_j(P),$$

$$\beta_{Face} = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Body} = \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Chair} = \frac{\sigma + \Sigma_k C_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Place} = \frac{\sigma + \Sigma_k P_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

where the coefficient of each category depends only on the ratio between the selectivity of the normalization pool for each of the categories and the selectivity for the other categories.

A face-selective neuron that resides in a face-selective area responds more to faces than each of the other categories, i.e.,

$$F_j \gg B_j, C_j, P_j$$

and is surrounded by neurons that are mostly face-selective, i.e.,

$$\sum_k F_k \gg \sum_k B_k,\ \sum_k C_k\ ,\ \sum_k P_k.$$

Based on the normalization equation, we can predict that the response to the multi-category scene will be dominated by the response to the face (Figure 1d):

$$\beta_{Face} > \beta_{Body},\ \beta_{Chair},\ \beta_{Room}$$

In addition, some category-selective areas reside in neighboring locations, and the border between them contains two populations of neighboring neurons that are selective for either one of the two categories – for example, an area with a similar proportion of neurons that are selective for faces and bodies but not for chairs and places, such as in the border between the face and body areas, i.e.,

$$\Sigma_k F_k \approx \Sigma_k B_k$$

$$\Sigma_k F_k, \Sigma_k B_k \gg \Sigma_k C_k, \Sigma_k P_k$$

Based on the normalization equation, we can predict that the response to the multi-category scene will be dominated by the responses both to the face and to the body (Figure 1f):

$$\beta_{Face}, \beta_{Body} > \beta_{Chair}, \beta_{Room}$$

$$\beta_{Face} \approx \beta_{Body}$$

In other words, in an area that is selective for faces and bodies (but not places or chairs), the response to the multi-category scene would be biased to the whole person, while filtering out the other categories presented in the multi-category scene, i.e., the chair and the room.

We can further see that the difference between the coefficients of two categories, for example, the face and the body, is given by:

$$\beta_{Face} - \beta_{Body} = \frac{\sigma + \sum_k F_k}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k} - \frac{\sigma + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$= \frac{\sum_k F_k - \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k},$$

i.e., the difference between the coefficients is determined by the difference in the selectivity of the normalization pool for the two categories. Thus, as shown before, for areas that contain proximal face- and body-selective clusters of neurons, the response to the multi-category scene is equally biased to the two categories (i.e., the face and the body), while the chair and the room are filtered out with coefficients close to zero,

$$\beta_{Chair}, \beta_{Room} \approx 0,$$

$$\Sigma_k C_k, \Sigma_k P_k \approx 0.$$

Finally, the sum of the coefficients is given by:

$$\beta_{Face} + \beta_{Body} + \beta_{Chair} + \beta_{Place} =$$

$$= \frac{4\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$= \frac{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k} + \frac{4\sigma}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$= 1 + \frac{4\sigma}{\sigma + \sum_k F_k + \sum_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

i.e., the sum of the coefficients is slightly higher than 1, being equal to 1 plus a small positive term. σ is usually a small positive number (Reynolds & Heeger, 2009).

When using fMRI to measure the response to a multi-category stimulus, we measure the BOLD signal, which is an estimate of the response of thousands of neurons in a small patch of cortex (e.g., a 2x2x2 mm^3 voxel). The response of all the neurons in a voxel can be written as:

$$\Sigma_j R_j(F + B + C + P) = \Sigma_j \left(\gamma \frac{F_j + B_j + C_j + P_j}{\sigma + \Sigma_{k(j)} F_{k(j)} + \Sigma_{k(j)} B_{k(j)} + \Sigma_{k(j)} C_{k(j)} + \Sigma_{k(j)} P_{k(j)}} \right),$$

where k(j) indicates the k'th neuron in the normalization pool of neuron j.

Assuming that all neurons in a given voxel have a similar normalization pool, i.e., a similar surrounding, we can rewrite the equation such that k is no longer a function of j:

$$\Sigma_j R_j(F + B + C + P) \approx \Sigma_j \left(\gamma \frac{F_j + B_j + C_j + P_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k} \right).$$

Now, following the exact same derivations made for a single neuron, we can rewrite the equation of the expected response as a linear combination of the sum of responses of the neurons with the same coefficients used for a single neuron:

$$\Sigma_j R_j(F + B + C + P) = \beta_{Face} \cdot \Sigma_j R_j(F) + \beta_{Body} \cdot \Sigma_j R_j(F) + \beta_{Chair} \cdot \Sigma_j R_j(C) + \beta_{Room} \cdot \Sigma_j R_j(P),$$

$$\beta_{Face} = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Body} = \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Chair} = \frac{\sigma + \Sigma_k C_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

$$\beta_{Room} = \frac{\sigma + \Sigma_k P_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k + \Sigma_k C_k + \Sigma_k P_k}$$

These coefficients are dependent on the local selectivity for each of the isolated categories, which can be measured effectively by fMRI.

Data Availability

Data that was collected in this study will be available at https://openneuro.org after publication. Tables of preprocessed data as well as the code that was used to generate the analysis, figures and statistics are available at https://github.com/gylab-TAU/Multi_Category_Scenes_fMRI_analysis.

Author Contributions

L.K. and G.Y. designed the experiments, interpreted the data, and wrote the paper. L.K. collected and analyzed the data.

References

Baeck, A., Wagemans, J., & de Beeck, H. P. (2013). The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule. *NeuroImage*, *70*, 37–47. https://doi.org/10.1016/j.neuroimage.2012.12.023

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2016). Human-object interactions are more than the sum of their parts. *Cerebral Cortex*, 1–13. https://doi.org/10.1093/cercor/bhw077

Bao, P., & Tsao, D. Y. (2018). Representation of multiple objects in macaque category-selective areas. *Nature Communications*, *9*(1), 1–16. https://doi.org/10.1038/s41467-018-04126-7

Beck, D. M., & Kastner, S. (2005). Stimulus context modulates competition in human extrastriate cortex. *Nature Neuroscience*, *8*(8), 1110–1116. https://doi.org/10.1038/nn1501

Bernstein, M., Oron, J., Sadeh, B., & Yovel, G. (2014). An Integrated Face–Body Representation in the Fusiform Gyrus but Not the Lateral Occipital Cortex. *Journal of Cognitive Neuroscience*, *26*(11), 2469–2478. https://doi.org/10.1162/jocn_a_00639

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*. https://doi.org/10.1163/156856897X00357

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., & Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms13995

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Desimone, R., & Duncan, J. (1995). Neural mechanism of selective visual attention. *Annu Rev Neurosci*, *18*, 193–222.

Downing, P. E., Bray, D., Rogers, J., & Childs, C. (2004). Bodies capture attention when nothing is expected. *Cognition*, *93*(1), 27–38. https://doi.org/10.1016/j.cognition.2003.10.010

Fisher, C., & Freiwald, W. A. (2015). Whole-agent selectivity within the macaque face-processing system. *Proceedings of the National Academy of Sciences*, *112*(47), 201512378. https://doi.org/10.1073/pnas.1512378112

Foster, C., Zhao, M., Bolkart, T., Black, M. J., Bartels, A., & Bülthoff, I. (2021). Separated and overlapping neural coding of face and body identity. *Human Brain Mapping*, *42*(13), 4242–4260. https://doi.org/10.1002/hbm.25544

Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange, N., Kennedy, D. N., Herbert, M. R., Bent, E. K., Koneru, V. K., Dieterich, M. E., Hodge, S. M., Rauch, S. L., Grant, P. E., Cohen, B. M., Seidman, L. J., Caviness, V. S., & Biederman, J. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, *162*(7), 1256–1265. https://doi.org/10.1176/appi.ajp.162.7.1256

Goldstein, J. M., Seidman, L. J., Makris, N., Ahern, T., O'Brien, L. M., Caviness, V. S., Kennedy, D. N., Faraone, S. V., & Tsuang, M. T. (2007). Hypothalamic Abnormalities in Schizophrenia:

Sex Effects and Genetic Vulnerability. *Biological Psychiatry*, *61*(8), 935–945. https://doi.org/10.1016/j.biopsych.2006.06.027

Harry, B. B., Umla-Runge, K., Lawrence, A. D., Graham, K. S., & Downing, P. E. (2016). Evidence for integrated visual face and body representations in the anterior temporal lobes. *Journal of Cognitive Neuroscience*, *28*(8), 1178–1193. https://doi.org/10.1162/jocn_a_00966

Heeger, D. J. (2011). Heeger - 1992 - Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2).

Hu, Y., Baragchizadeh, A., & O'Toole, A. J. (2020). Integrating faces and bodies: Psychological and neural perspectives on whole person perception. *Neuroscience and Biobehavioral Reviews*, *112*(October 2019), 472–486. https://doi.org/10.1016/j.neubiorev.2020.02.021

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.12.065

Kaiser, D., Strnad, L., Seidl, K. N., Kastner, S., & Peelen, M. V. (2014). Whole person-evoked fMRI activity patterns in human fusiform gyrus are accurately modeled by a linear combination of face- and body-evoked activity patterns. *Journal of Neurophysiology*, *111*(1), 82–90. https://doi.org/10.1152/jn.00371.2013

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? *Perception*. https://doi.org/10.1068/v070821

Kliger, L., & Yovel, G. (2020). The Functional Organization of High-Level Visual Cortex Determines the Representation of Complex Visual Stimuli. *The Journal of Neuroscience*, *40*(39), 7545–7558. https://doi.org/10.1523/JNEUROSCI.0446-20.2020

MacEvoy, S. P., & Epstein, R. a. (2009). Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Current Biology : CB*, *19*(11), 943–947. https://doi.org/10.1016/j.cub.2009.04.020

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

Makris, N., Goldstein, J. M., Kennedy, D., Hodge, S. M., Caviness, V. S., Faraone, S. V., Tsuang, M. T., & Seidman, L. J. (2006). Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia Research*, *83*(2–3), 155–171. https://doi.org/10.1016/j.schres.2005.11.020

Mayer, K. M., Vuong, Q. C., & Thornton, I. M. (2015). Do people 'pop out'? *PLoS ONE*, *10*(10), 1–15. https://doi.org/10.1371/journal.pone.0139618

McMains, S., & Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, *31*(2), 587–597. https://doi.org/10.1523/JNEUROSCI.3766-10.2011

Miller, E. K., Gochin, P. M., & Gross, C. G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research*, *616*(1–2). https://doi.org/10.1016/0006-8993(93)90187-R

Op de Beeck, H. P., Pillet, I., & Ritchie, J. B. (2019). Factors Determining Where Category-Selective Areas Emerge in Visual Cortex. *Trends in Cognitive Sciences*, *23*(9), 784–797. https://doi.org/10.1016/j.tics.2019.06.006

Pessoa, L., Kastner, S., & Ungerleider, L. G. (2003). Neuroimaging studies of attention: From modulation of sensory processing to top-down control. *Journal of Neuroscience*, *23*(10), 3990–3998. https://doi.org/10.1523/jneurosci.23-10-03990.2003

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative fMRI study. *Journal of Neurophysiology*, *101*(5), 2581–2600. https://doi.org/10.1152/jn.91198.2008

Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). *Representations of faces and body parts in macaque temporal cortex: A functional MRI study* (Vol. 102, Issue 19). www.pnas.orgcgidoi10.1073pnas.0502605102

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/J.NEUROIMAGE.2011.10.018

Premereur, E., Taubert, J., Janssen, P., Vogels, R., & Vanduffel, W. (2016). Effective Connectivity Reveals Largely Independent Parallel Networks of Face and Body Patches. *Current Biology*, *26*(24), 3269–3279. https://doi.org/10.1016/j.cub.2016.09.059

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Reddy, L., & Kanwisher, N. (2007). Category Selectivity in the Ventral Visual Pathway Confers Robustness to Clutter and Diverted Attention. *Current Biology*. https://doi.org/10.1016/j.cub.2007.10.043

Reddy, L., Kanwisher, N. G., & Vanrullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21447–21452.

Reynolds, J. H., & Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, *61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002

Rolls, E. T., & Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Experimental Brain Research*, *103*(3). https://doi.org/10.1007/BF00241500

Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D. E., & Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, *19*(9), 1250–1255. https://doi.org/10.1038/nn.4354

Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience*, *25*(47). https://doi.org/10.1523/JNEUROSCI.2621-05.2005

Song, Y., Luo, Y. L. L., Li, X., Xu, M., & Liu, J. (2013). Representation of Contextually Related Multiple Objects in the Human Ventral Visual Pathway. *Journal of Cognitive Neuroscience*, *25*(8), 1261–1269. https://doi.org/10.1162/jocn_a_00406

Stein, T., Sterzer, P., & Peelen, M. V. (2012). Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. *Cognition*, *125*(1), 64–79. https://doi.org/10.1016/j.cognition.2012.06.005

Taubert, J., Ritchie, J. B., Ungerleider, L. G., & Baker, C. I. (2022). One object, two networks? Assessing the relationship between the face and body-selective regions in the primate visual system. *Brain Structure and Function*, *227*(4), 1423–1438. https://doi.org/10.1007/s00429-021-02420-7

Thompson, J. C., Clarke, M., Stewart, T., & Puce, A. (2005). Configural processing of biological motion in human superior temporal sulcus. *Journal of Neuroscience*, *25*(39), 9059–9066. https://doi.org/10.1523/JNEUROSCI.2129-05.2005

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, *114*(22), E4501–E4510. https://doi.org/10.1073/pnas.1612862114

Wachsmuth, E., Oram, M. W., & Perrett, D. I. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, *4*(5), 509–522. https://doi.org/10.1093/cercor/4.5.509

Weiner, K. S., & Grill-Spector, K. (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *NeuroImage*, *52*(4). https://doi.org/10.1016/j.neuroimage.2010.04.262

Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. In *Psychological Research* (Vol. 77, Issue 1). https://doi.org/10.1007/s00426-011-0392-x

Yovel, G., & Freiwald, W. A. (2013). Face recognition systems in monkey and human: Are they the same thing? *F1000Prime Reports*. https://doi.org/10.12703/P5-10

Zafirova, Y., Bognár, A., & Vogels, R. (2024). Configuration-sensitive face-body interactions in primate visual cortex. *Progress in Neurobiology*, *232*(November 2023). https://doi.org/10.1016/j.pneurobio.2023.102545

Zafirova, Y., Cui, D., Raman, R., & Vogels, R. (2022). Keep the head in the right place: Face-body interactions in inferior temporal cortex. *NeuroImage*, *264*(September), 119676. https://doi.org/10.1016/j.neuroimage.2022.119676

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(36), 8150–8164. https://doi.org/10.1523/JNEUROSCI.2058-05.2005