Cerebral Cortex Advance Access published April 12, 2016

OXFORD

*Cerebral Cortex*, 2016, 1–13

**doi: 10.1093/cercor/bhw077**
Original Article

ORIGINAL ARTICLE

# Human–Object Interactions Are More than the Sum of Their Parts

Christopher Baldassano[1], Diane M. Beck[2] and Li Fei-Fei[1]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA and [2]Department of Psychology and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Address correspondence to Christopher Baldassano, Princeton Neuroscience Institute, Washington Road, Princeton, NJ 08544, USA.
Email: chrisb@princeton.edu



## Abstract

Understanding human–object interactions is critical for extracting meaning from everyday visual scenes and requires integrating complex relationships between human pose and object identity into a new percept. To understand how the brain builds these representations, we conducted 2 fMRI experiments in which subjects viewed humans interacting with objects, noninteracting human–object pairs, and isolated humans and objects. A number of visual regions process features of human–object interactions, including object identity information in the lateral occipital complex (LOC) and parahippocampal place area (PPA), and human pose information in the extrastriate body area (EBA) and posterior superior temporal sulcus (pSTS). Representations of human–object interactions in some regions, such as the posterior PPA (retinotopic maps PHC1 and PHC2) are well predicted by a simple linear combination of the response to object and pose information. Other regions, however, especially pSTS, exhibit representations for human–object interaction categories that are not predicted by their individual components, indicating that they encode human–object interactions as more than the sum of their parts. These results reveal the distributed networks underlying the emergent representation of human–object interactions necessary for social perception.

**Key words:** action perception, cross-decoding, fMRI, MVPA, scene perception

## Introduction

Our visual experience consists not of a jumble of isolated objects but of coherent scenes, in which objects are arranged in meaningful relationships. Neuroscientists have long studied recognition of isolated objects, and we have at least a qualitative understanding of where and how the brain constructs invariant object representations (DiCarlo et al. 2012). A largely separate body of research has studied the perception of complex scene images containing diverse collections of objects and has identified brain regions supporting the recognition of broad scene categories (Walther et al. 2009; MacEvoy and Epstein 2011). The connection between these 2 domains, however, has gone largely unstudied: how do objects come together to compose complex scenes with emergent semantic properties?

One scene category in which semantic meaning is critically driven by the relationship between scene components is that of human–object interactions. Understanding the differences between images of people riding horses, petting horses, leading horses and feeding horses, for example, cannot be accomplished by simply recognizing the person and horse in isolation. Moreover, understanding human–object interactions is essential for both developmental learning about object manipulation (Want and Harris 2002) as well as everyday social cooperation. Yet, we know surprisingly little about how such interactions are encoded in the brain.

It is clear that understanding human–object interactions will depend on brain regions involved in processing object identity (e.g., the lateral occipital complex, LOC) and the relative positions of body parts (e.g., the extrastriate body area, EBA). We hypothesize, however, that extracting meaning from human–object interactions will require areas sensitive not just to object or pose, but also to higher order emergent features of the interaction relevant

to understanding human's actions and goals. In other words, we expect that the object and human representations must be integrated into a neural representation that is "more than the sum of its parts."

To identify such a representation, we used multi-voxel pattern analysis (MVPA) to compare the representation of human–object interaction categories to the linear combinations of responses evoked by isolated humans and objects. Although some multiobject scenes can be modeled by a linear pattern average of the responses to each object individually (Zoccolan, Cox, and DiCarlo, 2005; MacEvoy and Epstein 2009; Baeck et al. 2013; Kaiser et al. 2014; Kubilius et al. 2015), we find that human–object interactions break this linear assumption in regions such as the posterior superior temporal sulcus (pSTS), evoking novel category representations distinct from pattern averages. In particular, this analysis revealed nonlinear representations across multiple components of the social cognition network (Saxe 2006).

We conclude that understanding human–object interactions involves distributed occipitotemporal networks, which support the creation of emergent representations in social cognition regions. These results demonstrate the critical impact of interactions between scene components on scene representation, providing a new bridge between isolated object perception and full scene recognition.

## Materials and Methods

### Subjects

We collected data from 10 subjects (2 female, aged 22–28, including one of the authors) in Experiment 1, and 12 subjects (5 female, aged 20–32, including one of the authors, 5 subjects overlapping with first experiment) in Experiment 2. Subjects were in good health with no past history of psychiatric or neurological diseases and with normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board of Stanford University, and all subjects gave their written informed consent.

### Stimuli

For Experiment 1, we created 128 person-riding-horse and 128 person-playing-guitar images by manually segmenting images from the Stanford 40 Actions database (Yao et al. 2011). Each image was scaled to contain the same number of pixels, such that every image fits with a 450 × 450 square. We created 128 horse images (using images and masks from the Weizmann horse database; Borenstein and Malik 2006) and 128 guitar images (using images from the Caltech Guitar dataset, and manually segmenting them from the background; Tirilly et al. 2008). We also created 128 person images using images and masks from INRIA Annotations for Graz-02 (Opelt et al. 2006; Marszalek and Schmid 2007) in additional to manually segmented people from the Stanford 40 Actions database (Yao et al. 2011). Each of the isolated images was scaled to contain half as many pixels as the interacting images. Half of the horses were horizontally mirrored (since all of the Weizmann horses face to the left), and the guitars were rotated so that the distribution of the neck angles exactly matched that of the person-playing-guitar images.

To create the noninteracting images, we overlaid an isolated person and isolated object, with the person and object chosen so as to avoid pairings that appeared to be interacting. The person and object images were each centered on a point drawn from a Gaussian distribution around the fixation point, with standard deviation set equal to the standard deviation of objects and people relative to the image centers in the action images (0.62° of visual angle). To make the images as qualitatively similar to the action images as possible, the person images were placed on top of (occluding) the horse images, but were placed behind the guitar images. The distribution of the relative sizes of the person and object was exactly matched to that of the action images, and the composite images were scaled to have the same number of pixels as the interacting images. The total number of stimuli in Experiment 1 was (3 isolated + 2 interacting + 2 noninteracting) × (128 images) = 896 images.

For Experiment 2, 40 images were collected from Google Images and Flickr for each of 4 action categories: pushing shopping carts, pulling luggage, using a computer, and using a typewriter. All of the 160 images were manually segmented to remove the person and object from the background, and scaled to have the same number of pixels such that every image fits within a 900 × 900 square. In addition to the segmented human–object interaction image, we manually separated the person and object, creating isolated object images and isolated human images. Any overlap between the person and object in the human–object interaction images was covered with a black rectangle (to ensure that the isolated person images did not contain any information about the object and vice versa), which was applied to all 3 versions of the image. All images were superimposed on a background containing $1/f$ noise in each color channel, in both their original orientation and mirrored left to right, for a total of (2 orientations) × (4 categories) × (3 conditions) × (40 images) = 960 stimuli.

### Experimental Design

Each subject viewed blocks of images from different categories, with a 12 s gap between blocks. Each block started with a 500 ms fixation cross, and then 8 images each presented for 160 ms with a 590 ms blank inter-trial interval. Subjects were instructed to maintain fixation at the center of the screen and perform a 1-back task using a button box. In Experiment 1, subjects participated in 8 runs, each of which contained 2 blocks of each of the 7 stimulus categories (isolated humans, guitars, and horses; noninteracting human–guitar and human–horse pairs; humans riding horses and humans playing guitars), for a total of 14 blocks (126 TRs) per run. Subjects performed a 1-back task, detecting consecutive repetitions of the same image, which occurred 0, 1, or 2 times per block. In the Experiment 2, subjects performed 14 runs. Each of the first 10 runs contained 8 blocks, one from every isolated (person/object) category, for a total of 79 TRs per run. The last 4 runs contained 20 blocks each (10 per run), with 5 blocks drawn from each interaction category, for a total of 97 TRs per run. Subjects performed a 1-back task, detecting consecutive images that were mirror images of each other, which occurred 0 or 1 times per block (with the same frequency for all categories and conditions). Note that in both experiments, every stimulus image was distinct, aside from the adjacent repeated images in Experiment 1 which were never separated in the analyses. Thus, all MVPA training and testing were done on distinct images.

### Regions of Interest

The locations of the category-selective ROIs for each subject's brain were obtained using standard localizer runs conducted in a separate fMRI experiment. Subjects performed 2 runs, each with 12 blocks drawn equally from 6 categories—child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects—and an additional run with 12 blocks drawn from 2 categories (body parts and

objects). Blocks were separated by 12 s fixation cross periods and consisted of 12 image presentations, each of which consisted of a 900 ms image followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of 2 images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen and respond via button press whenever an image was repeated. The ROIs were defined such that each subject had approximately the same total volume of clustered voxels: LOC, approximately 4800 $mm^3$ for Objects > Scrambled contrast in lateral occipital cortex; EBA, peak clusters of approximately 2900 $mm^3$ for Body Parts > Objects contrast in occipital cortex; parahippocampal place area (PPA), peak clusters of approximately 2900 $mm^3$ for Scenes > Objects contrast near the parahippocampal gyrus; fusiform face area (FFA), peak clusters of approximately 960 $mm^3$ for Faces > Objects contrast near the fusiform gyrus. The volume of each ROI in $mm^3$ was chosen conservatively, based on previous results (Golarai et al. 2007).

We defined retinotopic regions PHC1/2 using a group-level field map atlas (Wang et al. 2014).We also defined a pSTS ROI for Experiment 2 in MNI space as all voxels within 10 mm of the peak pSTS voxel in Experiment 1 (see Fig. 4). Both of these 2 ROIs were then transformed into each subject's native space.

## Scanning Parameters and Preprocessing

For Experiment 1 and the ROI localizers, imaging data were acquired with a 3 T G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 80°; matrix, 128 × 128 voxels; FOV, 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution, 1.56 × 1.56 mm]. The first 4 volumes of each run were discarded, and the functional data were then motion-corrected and converted to percent signal change, using the AFNI software package (Cox 1996). Since we are using MVPA, no other preprocessing was performed. We collected a high-resolution (1 × 1×1 mm voxels) structural scan (SPGR; TR, 5.9 ms; TE, 2.0 ms; flip angle, 11°) in each scanning session. For computing whole-brain results at the group level, each subject's anatomy was registered by hand to the Talaraich coordinate system. Images were presented using a back-projection system (Optoma Corporation) operating at a resolution of 1024 × 768 pixels at 75 Hz, such that images covered approximately 14° of visual angle.

For Experiment 2, imaging data were acquired with a different 3 T G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 77°; matrix, 80 × 80 voxels; FOV, 23.2 cm; 42 oblique 2.9 mm slices; in-plane resolution, 2.9 × 2.9 mm]. The first 6 volumes of each run were discarded, and the functional data were then motion-corrected and converted to percent signal change, using the AFNI software package (Cox 1996). We collected a high-resolution (0.9 × 0.9 × 0.9 mm voxels) structural scan (BRAVO; TR, 7.24 ms; TE, 2.78 ms; flip angle, 12°) in each scanning session. No other preprocessing was performed. For computing whole-brain results at the group level, each subject's anatomy was registered automatically to the MNI coordinate system. Images were presented using an LCD display (Resonance Technology) operating at a resolution of 640 × 480 at 240 Hz, visible from a mirror within the head coil, such that images covered approximately 12° of visual angle.

## Mean Signal Analysis

To compare the mean signal response to noninteracting and interacting stimuli in each ROI in Experiment 1, we used a standard regression model. The stimulus regressors were modeled as step functions equal to 1 during a stimulus block and 0 elsewhere, convolved with the standard AFNI hemodynamic response function (Cox 1996). In addition, 30 nuisance regressors were added to the model: 3 for each of the 8 runs (constant offset, linear trend, quadratic trend), and 6 motion correction estimates (3 rotation and 3 translation). The estimated β weights for the noninteracting and interacting regressors were then recorded in units of percent signal change relative to the mean of the run.

## ROI Decoding

For all MVPA decoding analyses in both Experiments, each fMRI time point was first assigned a stimulus label; all time points that occurred during a stimulus block (shifted by 6 seconds to account for hemodynamic lag) were assigned to the corresponding stimulus label, while all other time points were labeled as interblock time points. Classification was performed using linear support vector machines, using the MATLAB LIBSVM library (Chang and Lin 2011). In Experiment 1, we selected 6 runs for training, used 1 validation run to tune the soft-margin hyperparameter $c$, and tested on the remaining run. Results were averaged over all possible choices of testing and validation runs. In Experiment 2, 9 blocks of each stimulus category were selected for training, and the classifier was then tested on the remaining blocks, with fixed $c = 0.1$. Results were averaged over all choices of testing block. For cross-decoding, the classifier was also tested on all blocks corresponding to the untrained stimulus conditions.

When applying this method to the predefined ROIs, we first excluded voxels that were not sensitive to visual stimulation, to improve decoding accuracy. All voxels were ranked based on the absolute value of their z-score for within-block time points (i.e., visual stimulation) versus interblock time points (i.e., blank screen with fixation point). The top 40% of the voxels were used in decoding (the number of voxels retained was set to 40% of the group mean size for each region, so all subjects retained the same number of voxels in a given region), but our results are not sensitive to the number of voxels used (see Supplementary Fig. 1). Note that this type of voxel selection does not introduce a circularity bias (as described by Vul et al. 2009) since 1) we are selecting only for visual sensitivity, not for between-condition effects, and 2) the selection is based only on training data.

In Experiment 1, 2 separate classifiers were trained: one to discriminate between noninteracting stimulus categories (humans with horses vs. humans with guitars) and one to discriminate between interacting stimulus categories (humans riding horses vs. humans playing guitars). In the first analysis, the performance of these classifiers was measured on the noninteracting and interacting testing time points, respectively. For the cross-decoding analysis aimed at identifying nonlinear interactions of human and object, we created pattern-average testing time points, by averaging the mean response to isolated humans in the testing run with all isolated object time points in the testing run. The noninteracting and interacting decoders were then used to classify the category of these pattern average time points (human + horse vs. human + guitar) as well as the category of isolated object time points (horse vs. guitar). Because the noninteracting decoder could not have learned an emergent interaction we would expect it to transfer well (i.e., above chance cross-decoding) to both the pattern averages as well as the isolated objects. Similarly, above-chance accuracy for the interacting decoder in this cross-decoding analysis would indicate that the category representation of a human–object interaction is at least partially

predicted by the object alone, or from a linear average of the person and object representations. A reduction in accuracy relative to the noninteracting decoder, however, would suggest that the interacting decoder learned something beyond the component objects.

In Experiment 2, 3 classifiers were trained: one to discriminate between isolated objects segmented from our action images, one to discriminate between isolated humans segmented from our action images, and one to discriminate between the full human–object action images. This last classifier was also applied in a cross-decoding analysis, to decode isolated object time points, isolated human time points, and pattern-average time points (created by averaging the 4 time points corresponding to an isolated object category in a given run with the 4 time points corresponding to the isolated human from the same category in the same run, yielding a new set of 4 pattern-average time points). As in Experiment 1, above-chance accuracy for the full-interaction decoder on these isolated objects, humans and pattern-average time points would suggest that the representation for interactions shares similarities to isolated object category representations, isolated human pose category representations, or the average of the 2, respectively. A failure of the full-interaction decoder to transfer to either the isolated stimuli or pattern averages would suggest that the full-interaction decoder learned something beyond the components.

### MVPA Searchlight Analyses

As exploratory analyses, we also ran these analyses in a whole-brain searchlight. Spheres with 7 mm radius were centered on a grid with 8 mm spacing. For each sphere, all voxels whose centers fell within its radius were used as a region of interest, and decoding analyses were performed as for the ROIs (without any voxel selection, and with a soft-margin hyperparameter set to the average of its value during the ROI experiments). Note that each sphere intersected with all 26 neighboring spheres, since the maximum distance between a sphere and its neighbors (square root of 3 times 8) is less than twice the radius (2 times 7). To produce a decoding accuracy map for each subject, the accuracy for each voxel was calculated as the mean accuracy of all searchlights that included that voxel.

To determine significance thresholds, a Monte-Carlo permutation test was used (Stelzer et al. 2013). The analysis used on the real data was run 1000 times on data for which the time point labels were randomly shuffled between categories being used for training or testing. For example, when decoding riding horse versus playing guitar, the labels of all riding-horse and playing-guitar time points were randomly shuffled. A threshold value was then fixed such that <5% of the sampled maps contained any above-threshold clusters larger than 100 voxels, and this same threshold was applied to the real data (see Supplementary Fig. 2). This nonparametric correction procedure has been shown to be much more conservative than parametric statistical methods, which can highly inflate family-wise error rates (Eklund et al. 2015).

## Results

### Experiment 1

We constructed a stimulus set with 3 types of images (see Fig. 1): isolated humans, guitars, and horses; "noninteracting" human–horse and human–guitar pairs, in which humans and objects were simply pasted together without an interaction; and "interacting" humans riding horses and humans playing guitars. These actions were chosen since both involve humans that are roughly vertical and centered, so that the noninteracting and interacting images had similar construction. As described in Experimental Procedures, the noninteracting images were constructed to match the statistics of the interacting images as closely as possible, so that the only difference from the interacting images is that the human body is not correctly positioned to interact with the object.

### Category Decoding for Noninteracting and Interacting Stimuli

Not surprisingly, given the subtle differences in the stimuli, a univariate analysis comparing interacting and noninteracting stimuli yielded no differences in occipitotemporal regions LOC and EBA (LOC: $t_9 = -0.47$, $P = 0.65$; EBA: $t_9 = -0.78$, $P = 0.46$; 2-tailed $t$-test) and performing a whole-brain regression analysis contrasting interacting > noninteracting failed to find any voxels meeting the threshold of FDR < 0.05. Thus, we used MVPA decoding to find regions that showed different voxel-wise patterns between conditions. In particular, to identify regions that may be sensitive specifically to interactions, we looked for regions where we were better able to decode between the human–guitar and human–horse stimuli when they were interacting than when they were noninteracting. Such a result would indicate that a region better distinguishes between the 2 human–object categories when an interaction is present, implying that this region contains specialized processing for human–object interactions. We found that the category (horse vs. guitar) could be decoded for both noninteracting and interacting stimuli in all 3 of our regions of interest (noninteracting: LOC: $t_9 = 4.19$, $P = 0.001$; EBA: $t_9 = 3.24$, $P = 0.005$; interacting: LOC: $t_9 = 3.50$, $P = 0.003$; EBA: $t_9 = 5.41$, $P < 0.001$; 1-tailed $t$-test). LOC showed nearly identical decoding rates for both stimulus types ($t_9 = -0.13$, $P = 0.90$; 2-tailed $t$-test), but EBA showed a consistent difference in the decoding rates for noninteracting and interacting stimuli, with significantly better category decoding for interacting stimuli (EBA: $t_9 = 2.82$, $P = 0.020$; 2-tailed $t$-test). These results are shown in Figure 2 (solid bars, N → N and I → I). A searchlight analysis for areas showing this same preference for interacting stimuli (Fig. 3) produced areas consistent with our ROI results; we found voxels in right EBA that gave better decoding for interacting stimuli. Additionally, this contrast revealed a more anterior patch of cortex around the right pSTS showing the same preference for interacting stimuli. EBA and pSTS therefore exhibit sharper (more tightly clustered) responses to action categories when an interaction is present between the human and object.

### Cross-Decoding to Isolated Objects and Pattern Averages

As discussed above, perceiving human–object interactions requires a representation that is more than the sum of its parts. As shown in previous work, some regions' response to a pair of simultaneously presented stimuli is simply the average of the responses to the individual stimuli (Zoccolan, Cox, and DiCarlo, 2005; MacEvoy and Epstein 2009; Baeck et al. 2013; Kaiser et al. 2014; Kubilius et al. 2015). If a region is sensitive to human–object interactions, however, we would expect the region's response to an interacting human and object to not be simply the sum of its parts, but to be qualitatively different from a simple average of human and object. We hypothesize that regions specifically sensitive to human–object interactions should have specialized

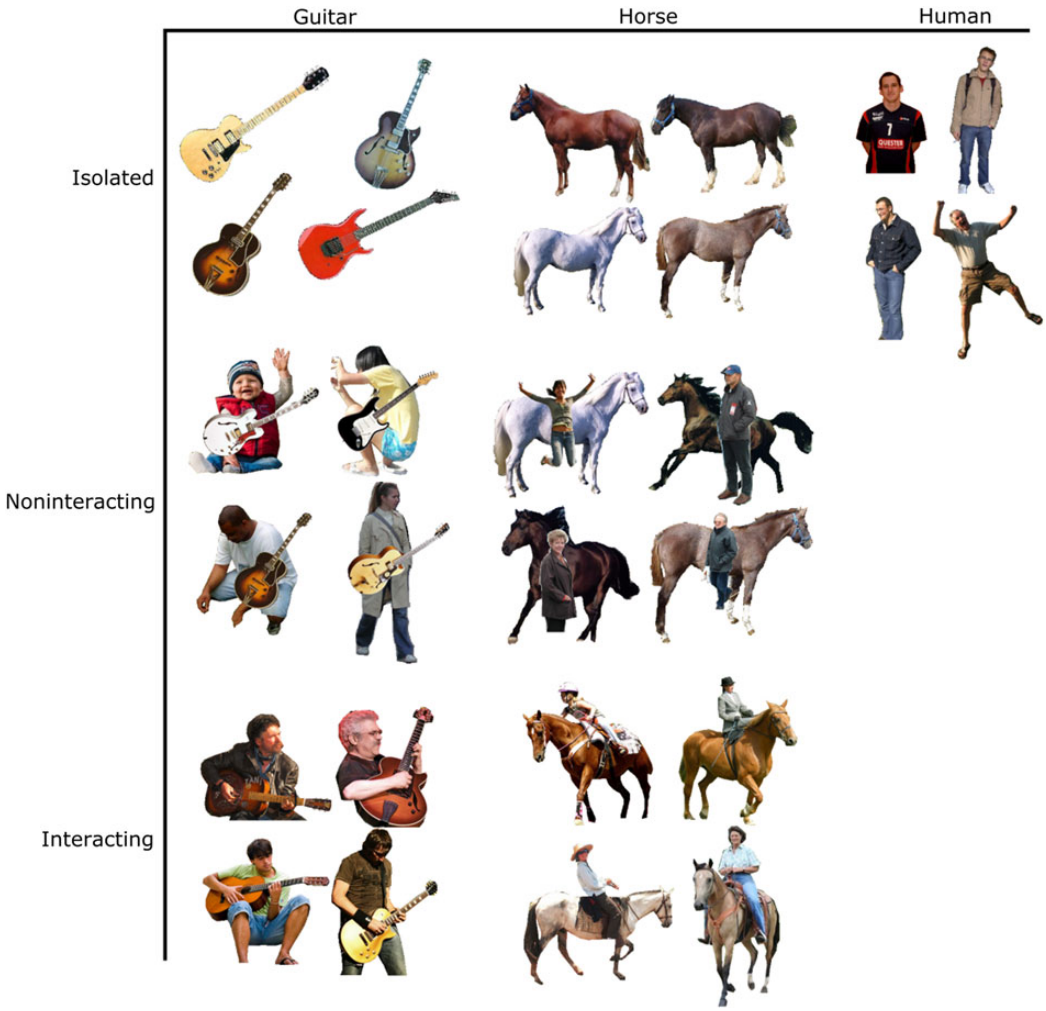


**Figure 1.** Example stimuli from Experiment 1. Subjects were shown 128 images in each of 7 categories: isolated guitars, horses, and people; noninteracting human–guitar pairs and human–horse pairs; and interacting humans playing guitars and humans riding horses.

(nonlinear) representations for categories of interacting human–object pairs, but not for noninteracting categories.

We can find regions showing this behavior by using a cross-decoding approach. After training 2 classifiers, as before, to decode noninteracting human–horses versus human–guitars, and to decode interacting human–horses versus human–guitars, we can then attempt to use these classifiers to decode isolated horses versus guitars as well as the pattern average of isolated humans and horses versus the pattern average of isolated humans and guitars. If the features used to represent categories of human–object pairs are driven simply by object information, or are simply linear averages of the features of isolated humans and objects, then both noninteracting and interacting classifiers trained on pairs should generalize well when cross-decoding. If, however, the classifier trained on interacting stimuli is decoding a representation that is more than the sum of its parts, then this classifier should be markedly worse at decoding patterns derived from responses to isolated stimuli than from responses to actual interacting stimuli.

The ROI results in Figure 2 show a compelling difference between cross-decoding in the noninteracting and interacting cases. When trained on noninteracting responses, classifiers for both regions were able to decode isolated object stimuli (N→O bars; LOC: $t_9 = 6.03$, $P < 0.001$; EBA: $t_9 = 5.27$, $P < 0.001$; 1-tailed $t$-test) and pattern-averaged stimuli (N→PA bars; LOC: $t_9 = 4.48$, $P < 0.001$; EBA: $t_9 = 4.72$, $P < 0.001$; 1-tailed $t$-test), with only a small drop in performance compared with decoding noninteracting stimuli (N→N > N→O: LOC: $t_9 = 0.02$, $P = 0.49$; EBA: $t_9 = 0.55$, $P = 0.29$; N→N > N→PA: LOC: $t_9 = 0.40$, $P = 0.34$; EBA: $t_9 = 0.94$, $P = 0.19$; 1-tailed $t$-test). This indicates that the features used to represent noninteracting stimulus categories are largely driven by object category information and can be effectively used to classify the average of the human and object patterns. Neither of these regions appears to represent noninteracting human–object pairs in a specialized, nonlinear way.

Cross-decoding results showed a different pattern, however, when the classifier was trained on interacting stimuli. In LOC, the interacting-stimulus classifier still showed some generalization to isolated objects ($t_9 = 2.44$, $P = 0.019$; 1-tailed $t$-test) and marginal performance on pattern-average responses ($t_9 = 1.78$, $P = 0.054$; 1-tailed $t$-test), with a nonsignificant drop compared with decoding interacting stimuli (objects: $t_9 = 0.91$, $P = 0.19$; pattern averages: $t_9 = 1.33$, $P = 0.11$; 1-tailed $t$-test). Thus, despite being trained on interacting stimuli, in LOC the classifier was still largely driven by the identity of the objects alone. In EBA, however, the classifiers trained on interacting stimuli showed a significant drop in performance when used to decode isolated objects ($t_9 = 3.09$, $P = 0.006$; 1-tailed $t$-test) or pattern averages

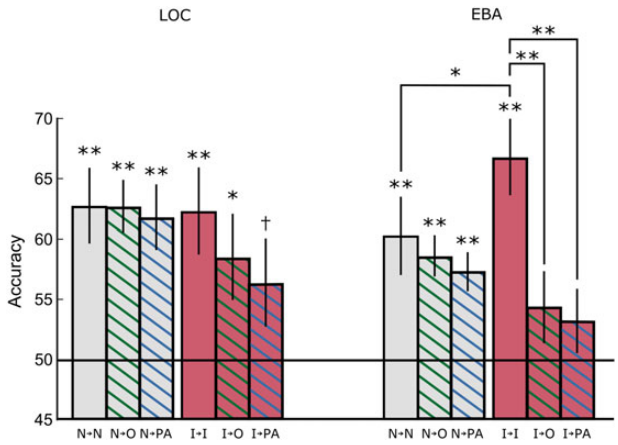


**Figure 2.** MVPA decoding and cross-decoding for Experiment 1. The stimulus category for images of human–object pairs (person and horse vs. person and guitar) can be decoded in both LOC and EBA, whether an interaction is present (I) or not (N). However, only EBA shows a significant increase in decoding accuracy for interacting stimuli (I→I) compared with noninteracting (N→N), indicating that the image category is better represented in this region when an interaction is present. Classifiers trained on responses to noninteracting stimuli in all 3 areas generalize well to isolated object responses (N→O) or pattern averages of individual humans and objects (N→PA), in both regions, indicating that category representations of noninteracting human–object pairs are linearly related to isolated human and object responses. The classifier trained on interacting humans and objects however, only generalizes marginally objects (I→O) to pattern averages (I→PA) in LOC and is near chance in EBA. This indicates that representation for human–object interaction categories, especially in EBA, cannot be captured by the average of responses to isolated humans (with uncontrolled poses) and objects. These results are consistent regardless of the number of voxels selected per region (see Supplementary Fig. 1). Error bars denote S.E.M., $^{\dagger}P = 0.054$, $^{*}P < 0.05$, $^{**}P < 0.01$.

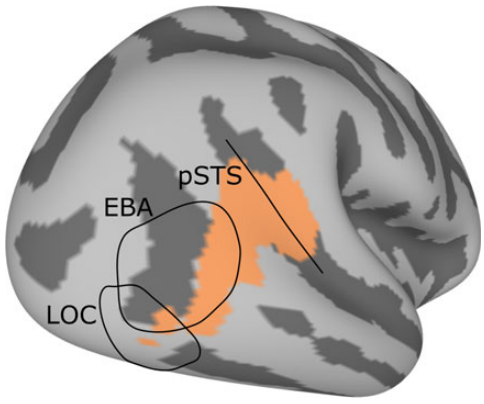


**Figure 3.** MVPA decoding difference searchlight for Experiment 1. Searching all of cortex for regions having higher decoding accuracy for interacting (I→I) than noninteracting (N→N) stimuli yields a result consistent with the ROI-based analysis. Searchlights showing this preference for interacting stimuli consistently included voxels in the anterior EBA and posterior STS in the right hemisphere. $P < 0.05$ cluster-level corrected.

($t_9 = 3.31$, $P = 0.005$; 1-tailed $t$-test) and were unable to decode the object or pattern-averaged stimuli above chance (objects: $t_9 = 1.49$, $P = 0.085$; pattern averages: $t_9 = 1.23$, $P = 0.12$; 1-tailed $t$-test). This drop was significantly larger than that in early visual areas ($t_9 = 2.85$, $P = 0.019$; 2-tailed $t$-test; see Supplementary Fig. 3), suggesting that it is being driven by more than simple visual dissimilarity between the isolated and interacting stimuli. In EBA, then, we have evidence that the classifier trained on interacting stimuli learnt a distinction that did not depend simply on the presence of the human and the object, suggesting that it was sensitive to either the interaction itself or the specific pose of the human in the interaction. We found a pattern in FFA similar to that in EBA, while we did not observe any significant decoding in PPA for these stimuli (see Supplementary Fig. 3).

Using a the same logic, we can look outside our ROIs and search for all regions with this pattern of results for human–object interaction categories by performing a searchlight analysis, identifying searchlights with a greater nonlinearity (drop in performance when cross-decoding) for interacting stimuli than noninteracting stimuli (Fig. 4). In addition to EBA, this contrast revealed regions around the pSTS (peak voxel at MNI [54, −43, 12]) and temporoparietal junction (TPJ) in both hemispheres, right dorsal PCC, and the right angular gyrus in the inferior parietal lobule (IPL) as decoding more than the sum of the parts. These areas largely map onto the network of regions involved in social cognition and understanding action intent (Saxe 2006), consistent with interactions between the human and object being an important component of the semantic meaning of a social scene. These results indicate that the representation of human–object interaction categories in these body-related regions is not simply driven by a linear combination of isolated object identity and a (nonpose specific) human activity pattern.

## Experiment 2

The results of Experiment 1 demonstrate that body-related regions do not represent "person riding horse" as a linear combination of "person" and "horse," but it is possible that some of this effect is due to differences in pose; although pose is in a sense a configural property of the human, pose representations do not incorporate both human and object information into a single emergent unit. In other words, we may not be decoding the interaction per se, but the fact that the interactions result in a particular pose. We tested this possibility using a new experiment that focused specifically on cross-decoding for interacting images, with a new larger set of stimuli (Fig. 5). Subjects viewed 4 new action categories, but also viewed in isolation the identical objects and humans extracted from these interaction images. This design ensured that objects and human poses were exactly matched between the isolated and interacting images, so that a failure to generalize decoding from interacting to pattern averaged responses would necessarily indicate a nonlinearity in category representation of interaction. In addition, we added a noise background behind each stimulus, to remove low-level object contour information and better simulate natural perception in a full visual scene.

We performed MVPA decoding using the same approach as in Experiment 1, looking now at 4-way action classification for objects alone, people alone, and person–object interactions. As before, we measured whether the representation of human–object interactions was similar to the representation of its components using cross-decoding; we applied the classifier trained on full interactions to classify objects alone, people alone, and pattern averages of objects and people. In addition to the ROIs used in Experiment 1 (LOC and EBA), we also defined a pSTS ROI with a 10 mm radius around the voxel that showed the strongest effect in Experiment 1 (Fig. 4).

The decoding results are displayed in Figure 6. Both LOC and EBA show above-chance decoding for objects, human poses, and interactions (objects: LOC $t_{11} = 6.12$, $P < 0.001$; EBA $t_{11} = 2.09$, $P = 0.030$; humans: LOC $t_{11} = 4.84$, $P < 0.001$; EBA $t_{11} = 2.30$, $P = 0.021$; interactions: LOC $t_{11} = 4.32$, $P < 0.001$; EBA $t_{11} = 2.93$, $P = 0.007$;


**Figure 4.** MVPA cross-decoding searchlight for Experiment 1. Colored voxels are those showing a larger nonlinearity in the interacting condition (I→I minus I→PA) compared with the nonlinearity in the noninteracting condition (N→N minus N→PA). In addition to EBA, this measure identifies regions around the posterior STS (peak effect marked with a dot) and TPJ in both hemispheres, the right dorsal PCC, and the right angular gyrus, $P < 0.05$ cluster-corrected.

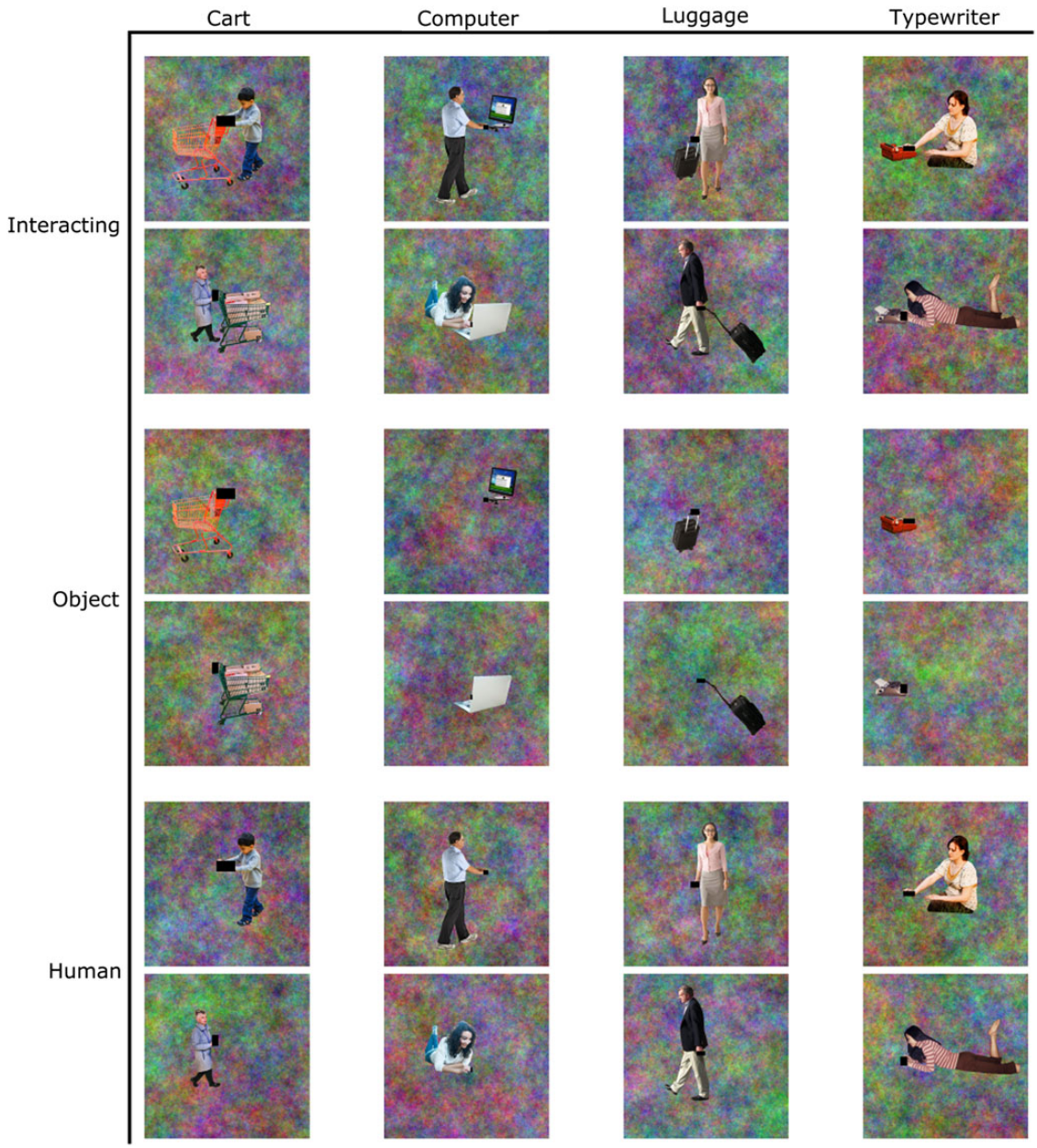


**Figure 5.** Example stimuli from Experiment 2. Subjects viewed images of human–object interactions from 4 different action categories (pushing carts, using computers, pulling luggage, and typing on typewriters) and also viewed the objects and people from these images in isolation.

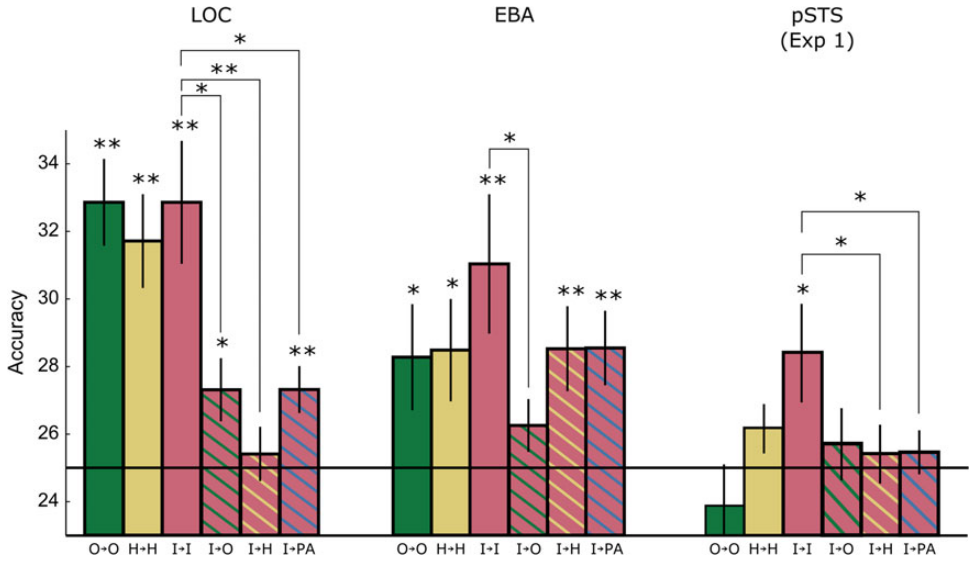


**Figure 6.** MVPA decoding and cross-decoding for Experiment 2. Both LOC and EBA show significant decoding of action category from isolated objects (green), isolated humans (yellow), or full actions (pink). As in Experiment 1, the classifier trained on full interactions performs above-chance on objects only in LOC, though the cross-decoding (striped bars) accuracy drop here is significant in both LOC and EBA. EBA's interaction classifier does, however, generalize well to human poses (while LOC's does not). When tested on pattern averages that now include class-specific pose information (unlike Experiment 1), both LOC and EBA classifiers show above-chance generalization, driven by object information in LOC and by pose information in EBA. The pSTS, on the other hand, localized based on results in Experiment 1, shows above-chance decoding only for human–object interactions and does not generalize to pattern averages. Error bars denote S.E.M., $^{*}P < 0.05$, $^{**}P < 0.01$.

1-tailed *t*-test). We can gain insight into what features are being used by the interacting classifier through cross-decoding. We first investigate whether information about the objects alone or human pose alone is being used for classification by testing whether the interacting classifier generalizes to decode the category of isolated objects or humans.

### Cross-Decoding to Isolated Objects in LOC and EBA

When decoding isolated object identity using the interacting decoder, LOC performs above-chance while EBA does not (LOC $t_{11} = 2.48$, $P = 0.015$; EBA $t_{11} = 0.068$, 1-tailed *t*-test), suggesting that LOC is using object identity information to classify interacting images whereas EBA is not. However, classification in LOC is not only relying on object identity for interaction classification since, like EBA, it shows a significant performance drop (LOC $t_{11} = 2.54$, $P = 0.014$; EBA $t_{11} = 2.35$, $P = 0.019$; 1-tailed *t*-test) when tested on isolated objects compared with the interacting human and object. One possibility for this pattern of results is that in LOC the interacting classifier can use pose to augment discrimination among categories, increasing its chances of successful classification relative to the object alone; indeed, the fact that both isolated poses and objects were successfully decoded in LOC is consistent with this idea. However, the fact that we see a similar drop in performance for the pattern average in LOC (see "Cross-decoding to pattern averages in LOC and EBA") suggests instead that LOC's representation of interactions is affected in a nonlinear way by the presence of the human as another object; adding the pose back in does not improve performance.

### Cross-Decoding to Isolated Human Poses in LOC and EBA

When cross-decoding isolated human poses using the interacting classifier, we find significant decoding only in EBA (LOC $t_{11} = 0.52$, $P = 0.31$, EBA $t_{11} = 2.81$, $P = 0.008$), indicating that pose information plays a substantial role in EBA's representation of interactions but not LOC's. Although the interacting decoder in both LOC and EBA shows a numeric drop in accuracy when tested on isolated poses, this did not reach significance in EBA (LOC $t_{11} = 4.22$, $P < 0.001$; EBA $t_{11} = 1.35$, $P = 0.10$; 1-tailed *t*-test), again consistent with pose information playing substantial role in EBA's representation of interactions.

Running a repeated-measures ANOVA with ROI (LOC, EBA) and testing set (objects, human poses) as factors yielded a significant interaction ($F_{1,11} = 16.87$, $P = 0.002$), indicating that the absence of the preferred category (objects in LOC and human pose in EBA) is more detrimental to cross-decoding performance than the absence of the nonpreferred category. Together, these results suggest that, in keeping with known category preferences, representations of human–objects are driven primarily by object identity in LOC (with humans being 1 class of objects) and primarily by pose information in EBA.

In summary, although the interaction decoders in LOC and EBA show little relationship to action category information in the nonpreferred stimulus, they still show a drop in accuracy when tested on their preferred stimulus in isolation relative to their performance on the full interaction. This suggests that, in both regions, category information from the preferred stimulus is modified in some nonlinear way when the stimulus appears in context with the nonpreferred stimulus.

### Cross-Decoding to Pattern Averages in LOC and EBA

We looked explicitly for this type of nonlinear interaction by testing the interacting decoder on a linear pattern average of the responses to the 2 isolated stimuli. This resulted in above chance classification in both LOC and EBA (LOC $t_{11} = 3.34$, $P = 0.003$; EBA $t_{11} = 3.22$, $P = 0.004$; 1-tailed *t*-test), confirming that both LOC and EBA encode interacting categories based at least partially on the individual components (object identity and human pose). Importantly, however, in LOC we see a significant drop in accuracy for the pattern averages relative to the full interacting images ($t_{11} = 2.72$, $P = 0.010$; 1-tailed *t*-test). Despite containing both components of the interaction, the pattern average is not sufficient to

decode at the level of a full interaction, presumably because it lacks the emergent features present in the original action image. Moreover, the addition of an appropriately positioned person without the emergent interaction provided no benefit relative to the isolated objects alone (objects vs. pattern averages $t_{11} = 0.01$, $P = 0.99$; 2-tailed $t$-test). Together, these results suggest that the interacting decoder is sensitive to something beyond a linear sum of a particular object and pose in LOC. In EBA, an analogous pattern is observed, but the drop in accuracy from the interacting classifier (I→I) to the pattern averages (I→PA) did not reach significance ($t_{11} = 1.23$, $P = 0.12$). The searchlight analysis (described in "Searchlight analyses") may explain why this drop failed to reach significance.

## Decoding and Cross-Decoding in pSTS

A qualitatively different pattern of results from both LOC and EBA was seen in the pSTS, just anterior to EBA (as identified in Experiment 1). Here, the only condition that decoded above chance with the interacting classifier was the full interactions (I→I; $t_{11} = 2.32$, $P = 0.020$; 1-tailed $t$-test); isolated object and pose decoding was not significant (objects: $t_{11} = -0.93$, $P = 0.81$; humans: $t_{11} = 1.57$, $P = 0.072$; 1-tailed $t$-test). The interaction decoder did not generalize to isolated objects, human poses, or the pattern average of the 2; none of these conditions were above chance (objects: $t_{11} = 0.64$, $P = 0.27$; humans: $t_{11} = 0.45$, $P = 0.33$; pattern averages: $t_{11} = 0.69$, $P = 0.25$; 1-tailed $t$-test), and the human and pattern average cross-decoding accuracies showed significant drops from pure interaction decoding (objects: $t_{11} = 1.23$, $P = 0.12$; humans: $t_{11} = 2.01$, $P = 0.035$; pattern averages: $t_{11} = 1.88$, $P = 0.043$; 1-tailed $t$-test). These results provide a strong confirmation of the conclusions of Experiment 1, implicating pSTS in representing full human–object interactions as more than the sum of their parts. See Supplementary Figure 3 for analyses of additional ROIs, showing that category representations in early visual cortex are largely explained by a linear pattern average while FFA shows results more similar to those in pSTS. These results also hold for weighted pattern averages of objects and human poses, which do not increase cross-decoding accuracy compared with an equal weighting (see Supplementary Fig. 4).

## Searchlight Analyses

To further investigate the posterior-to-anterior decoding differences in lateral temporal cortex, we performed an exploratory searchlight analysis to measure both interaction classification and generalization to pattern averages. As shown in Figure 7, the results were largely consistent with the ROI analyses;

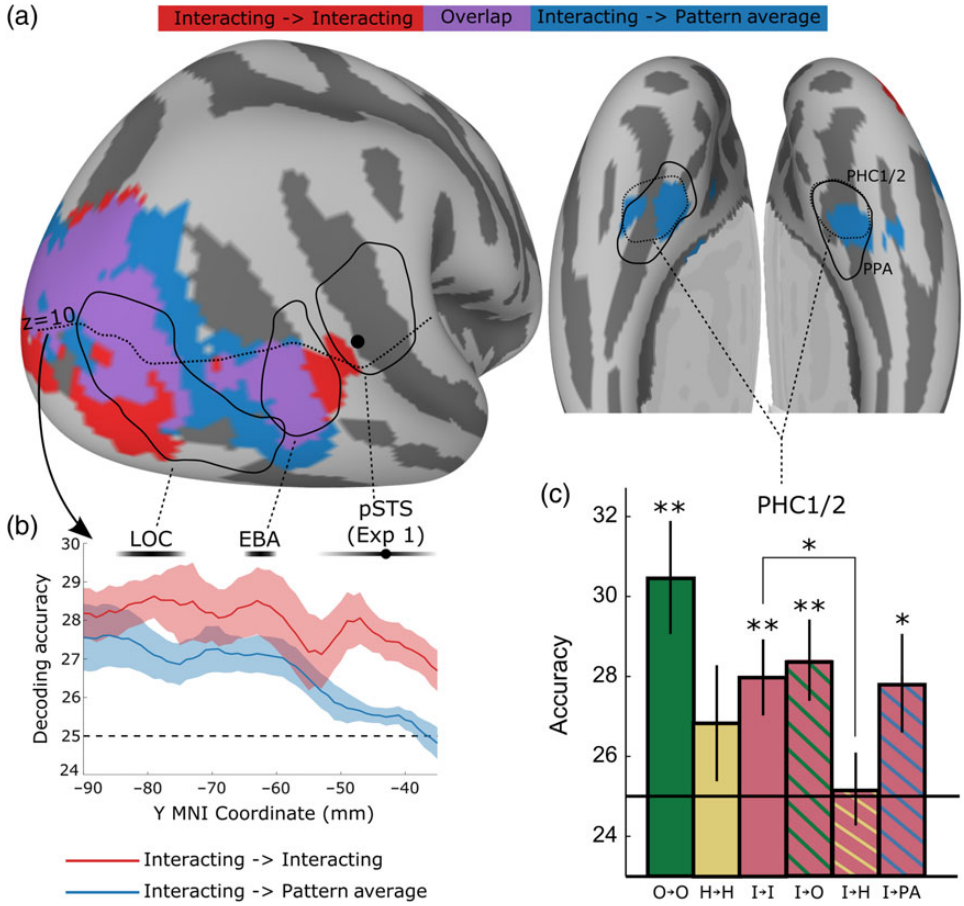


**Figure 7.** MVPA cross-decoding searchlight for Experiment 2. (*a*) As in Figure 6, we identified voxels that could decode the action category of human–object interactions and/or generalize this decoder to pattern averages. A large swath of right lateral occipital and temporal regions (including LOC and EBA) can classify interaction time points, but in only some portions of LOC and EBA (superior LOC and posterior EBA) does this classifier generalize to pattern averages. Regions in red (where the interacting classifier performs above chance, but fails to generalize to pattern averages) are likely candidates for processing emergent features of interactions. We also found significant generalization to pattern averages within the retinotopic (PHC1/2) regions of PPA. (*b*) A $z = 10$ slice of lateral cortex shows a clear difference between LOC/EBA and pSTS, with generalization to pattern averages much lower in pSTS. Error bars denote S.E.M. (*c*) Posterior PPA (PHC1/2) can decode both objects and interactions, and the interaction classifier generalizes fully to isolated objects (and pattern averages), indicating that this subregion is not highly sensitive to interactions.

classifying interactions was above chance in the majority of voxels within LOC and EBA, and each contained subregions (superior LOC and posterior EBA) where this classifier also generalized to decode pattern averages. In the anterior portion of EBA and pSTS, however, interaction cross-decoding fails on pattern averages; this posterior–anterior difference can be seen on an axial slice through lateral cortex (Fig. 7*b*), showing that cross-decoding accuracy drops rapidly around pSTS while interaction decoding remains relatively high. This trend suggests that the representation of interaction categories become less and less similar to pattern averages of humans and objects as we move anteriorly from LOC to EBA to pSTS. This result also explains why the EBA ROI failed to produce a significant drop in accuracy to the pattern averages; EBA contains a representation of interactions that is well predicted by a pattern average as well as a more anterior representation that represents the interaction as more than the sum of its parts.

### PHC1/2 Decoding

Interestingly, the searchlight also revealed significant cross-decoding in the PPA, restricted primarily to the retinotopic maps within this area (PHC1/2) (Arcaro 2009). We performed a post hoc ROI analysis of (independently defined) PHC1/2, which revealed a pattern of results different from all of the previous ROIs (Fig. 7*c*). The PHC maps represent object categories but not human pose categories (objects: $t_{11} = 3.95$, $P = 0.001$; humans: $t_{11} = 0.11$; 1-tailed *t*-test). Full interactions can be successfully decoded ($t_{11} = 3.23$, $P = 0.004$; 1-tailed *t*-test), and, critically, the interaction decoder generalizes to decode isolated objects or pattern averages, but not isolated humans (objects: $t_{11} = 3.45$, $P = 0.003$; humans: $t_{11} = 0.20$, $P = 0.42$; pattern averages: $t_{11} = 2.35$, $P = 0.019$; 1-tailed *t*-test). Unlike the other ROIs, there is no significant drop from interaction decoding to cross-decoding on isolated objects or pattern averages, but only for cross-decoding on isolated humans (objects: $t_{11} = -0.38$, $P = 0.65$; humans: $t_{11} = 2.36$, $P = 0.019$; pattern averages: $t_{11} = 0.11$, $P = 0.46$; 1-tailed *t*-test). The interacting category decoding is therefore largely explained by the isolated object representations, suggesting that emergent interaction features do not play a dominant role in action representations in this region. This indicates that the posterior portion of PPA may be driven primarily by the component objects of a scene rather than their relationships, at least for these simple 2-object scenes devoid of a background layout.

## Discussion

Using carefully constructed images of humans and objects, along with MVPA decoding and cross-decoding analyses, we identified regions in occipitotemporal cortex responsible for representing human pose and object identity, and for binding humans and objects together into a coherent interaction. Previous work has studied humans and objects in isolation (Downing and Peelen 2011; Kravitz et al. 2013), but we have characterized for the first time how categories of pose and object identity are jointly encoded in the context of human–object interactions.

### Lateral Occipital Complex

Decoding results in LOC revealed robust representations about action categories (i.e., humans interacting with objects), which were at least partially driven by object identity information (Experiments 1 and 2); that is, in both experiments a classifier trained on the interaction of object and human successfully generalized to isolated objects. Experiment 2 also suggested that LOC's representation of human–object interactions was not linearly related to the human's pose, since the interaction classifier was unable to classify isolated human poses above chance. However, LOC's responses to human–object interactions were partially driven by the emergent relationship between the human and object, since the interaction classifier showed a drop in performance (compared with full interactions) when given only isolated object or pattern averages. This failure to fully generalize to pattern averages was also seen in Experiment 1, but was only marginally significant; we note, however, that Experiment 1 had less data (fewer subjects and less data per subject) than Experiment 2, as well as other differences in stimuli and image acquisition. Taken together, these experiments show that LOC primarily represents object information in human–object interactions, but, at least for some action categories, may also be sensitive to features of the interaction between the human and object (especially in inferior LOC; see Fig. 7*a*).

### Extrastriate Body Area

EBA also showed consistent interaction decoding, but was not driven by object identity information (Experiments 1 and 2) and showed a similarity to pattern-averaged responses only when pose was carefully controlled (Experiment 2).

These results extend our current understanding of the role of EBA in action perception. It is well established that EBA represents body pose (reviewed in Downing and Peelen 2011). EBA, including the middle temporal gyrus (the most anterior portion of EBA, see Weiner and Grill-Spector 2011), has been implicated in action categorization through adaptation studies (Kable and Chatterjee 2006; Wiggett and Downing 2011), lesion studies (Kalénine et al. 2010), and a meta-analysis of object-related actions (Caspers et al. 2010). Exactly what type of information is represented in EBA has been more controversial, with proposals ranging from a "cognitively unelaborated" pose description (Downing and Peelen 2011) focused on "observable mechanics" (Spunt et al. 2010) to an amodal hub for pairing gestures with semantic meaning (Xu et al. 2009). Our results confirm that the EBA response to typical interactions is driven primarily by body pose (Experiment 2). However, the fact that noninteracting stimuli can be decoded above chance in Experiment 1 shows that EBA can discriminate based on object identity when the positioning of the human body is uninformative about the stimulus category. In addition, the decoder trained on full interactions failed to predict pattern average responses in anterior EBA, suggesting that at least portions of EBA could be sensitive to nonlinear relationships between the human and object. The fact that both object and pose information can be used by EBA raises the possibility that the representation in this region does represent more than simply body pose, though further work will be required to identify precisely how visual versus semantic this representation is.

### Posterior Superior Temporal Sulcus

The most interesting decoding trends with respect to the emergent properties of human–object interactions were observed in pSTS, which constructed representations of action categories that appear unrelated to object or pose information in isolation (Experiments 1 and 2). Overall, these results suggest that social cognition regions such as pSTS represent human–object interaction categories using specialized features that are not present in the linear averages of human and object patterns, creating representations of human–object interactions that are more than the sum of their parts.

The pSTS (and adjacent TPJ) regions anterior to EBA have been associated with more abstract types of action perception, such as understanding unusual or deceptive human action (Grézes et al. 2004; Brass et al. 2007) recognizing whether an object is being grasped in a typical way (Yoon et al. 2012) and many other tasks involving perception of agency, theory of mind, and Gestalt integration (Saxe and Kanwisher 2003; Pelphrey et al. 2004; Saxe et al. 2004; Decety and Lamm 2007; Hein and Knight 2008; Huberle and Karnath 2012). The pSTS has been proposed as the key hub for social perception, given its robust selectivity for many kinds of social content (Lahnakoski et al. 2012). Interestingly, although pSTS shows little sensitivity to object identity or pose (Experiment 2), we found specialized representations for interacting stimuli here in both Experiments. Therefore, pSTS appears to be less related to individual human or object representation, and more involved in understanding the visual or semantic features of full interactions.

### The Neural Basis of Action Recognition

There has been extensive prior work on the neural correlates of action perception, which is typically studied using video clips rather than controlled images (reviewed in Culham and Valyear 2006; Caspers et al. 2010). One controversy over the mechanism of action recognition is whether action recognition is carried out primarily in motor regions or in social reasoning areas. Under the simulation hypothesis, human actions are understood by mentally simulating the observed motor actions of the target and then inferring what the goals of the target must have been, a process presumed to be carried out in mirror neurons (Buccino et al. 2001; Buccino, Binkofski *et al.*, 2004; Buccino, Lui et al. 2004; Calvo-Merino et al. 2004, 2006; Rizzolatti and Craighero 2004; Chong et al. 2008). Under the teleological hypothesis, actions are understood by a more abstract social reasoning system, which does not depend on any mechanical "resonance" between the observer and target (Brass et al. 2007; Csibra 2007; Hickok 2009; Hauser and Wood 2010). Proponents of this view argue the activity seen in motor regions during action observation is involved in action prediction rather than action understanding (Csibra 2007; Lingnau et al. 2009) and that the type of errors made by action observers is inconsistent with mirror simulation theories (Saxe 2005).

Since our searchlight experiments show interaction effects almost exclusively in the social network (EBA, pSTS/TPJ, and PCC; Saxe 2006) and show no effects in motor or premotor cortex, our results provide support for the view that action understanding is built in social cognition regions, not in motor regions (Wheatley et al. 2007). Additionally, our data reveal that social cognition regions process action stimuli even in the absence of any social task, since our subjects were only performing 1-back repetition detection. However, it is possible that we do not observe motor or premotor involvement, because static images are not as effective at evoking a motor simulation. There was 1 region outside the social network identified by our study, the right IPL (in Experiment 1), which has been previously linked with action perception but whose precise function is unclear. Some work has argued that this region contains mirror neurons due to its cross-adaptation properties (Chong et al. 2008), but the stimuli that activate this region do not activate macaque mirror neurons (Hickok 2009) and lesion studies suggest that IPL is involved in the spatial coding of object-related actions, but not actual semantic action understanding (Goldenberg 2009; Kalénine et al. 2010).

### Comparison to Object–Object Interaction Studies

Previous work has attempted to link the perceptual grouping of interacting objects (Riddoch et al. 2003; Green and Hummel 2006; Roberts and Humphreys 2011) with activity in LOC, but the results have been controversial. Two studies have shown increased BOLD activity in LOC when objects are interacting (Kim and Biederman 2011) or positioned for interaction (Roberts and Humphreys 2010), while MVPA analyses have shown that the LOC response pattern for coherent scenes can be at least partially predicted as the average of responses to signature objects (MacEvoy and Epstein 2011) and that the LOC response to pairs of action-oriented objects is similar to a linear combination of the 2 object responses (Baeck et al. 2013).

Our results suggest that both camps are correct. LOC did not show a decoding preference for interaction versus noninteracting categorization (Experiment 1), has interaction representations that are at least partially related to isolated object identity (Experiments 1 and 2), and does not incorporate human pose information (Experiment 2). However, it does appear to encode some interaction information beyond object identity (Experiment 2), at least in the more inferior portion of LOC (red region in Fig. 7*a*). In other words, it is possible that the representation in LOC is modulated in some way by interactions, while still being primarily driven by linear combinations of isolated object identity information.

Our finding of significant cross-decoding in PPA is surprising, in light of previous work showing that PPA generates scene representations that are not predictable from the constituent objects of the scene (MacEvoy and Epstein 2011). There are several possible ways of reconciling these results. One possibility is that PPA's global scene representation applies to full photographic images, but not 2-element interactions, indicating that more complex stimuli are required to activate global processing in PPA (with more interactions or more explicit 3D geometry). Alternatively, global representations may be generated only in the anterior portion of PPA, while the posterior PHC1/2 subregion of PPA (Baldassano et al. 2015) accumulates local visual features in a way that is more similar to LOC (Baldassano et al. 2013).

### Identifying Configural Processing

Our approach for identifying regions sensitive to a relationship between stimulus features is a general tool that could be used to investigate other types of configural processing. For example, placing walls and a floor together to form a 3D room likely evokes a novel representation in regions sensitive to scene layout and navigation. Our analysis suggests that these regions would exhibit a large cross-decoding penalty when training on rooms and testing on the average response to walls and floors. The regions responsible for processing relational or contextual interactions between objects (Biederman et al. 1982) could be detected in a similar way, thereby avoiding the ambiguity of changes in mean BOLD activity. This approach for detecting configural representations gives researchers a new way to locate the areas critical for creating our rich, complex experience of the visual world.

## Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

## Funding

This work was funded by National Institutes of Health Grant 1 R01 EY019429 (to L.F.-F. and D.M.B.) and a National Science Foundation Graduate Research Fellowship under Grant No. DGE-0645962 (to C.B.).

## Notes

We thank the Richard M. Lucas Center for Imaging, the Center for Cognitive and Neurobiological Imaging, and the Stanford Vision lab for their comments and suggestions. *Conflict of Interest*: None declared.

## References

Arcaro MM. 2009. Retinotopic organization of human ventral visual cortex. J Neurosci. 29(34):10638–10652.

Baeck A, Wagemans J, Op de Beeck H. 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.

Baldassano C, Beck DM, Fei-Fei L. 2013. Differential connectivity within the parahippocampal place area. NeuroImage. 75:228–237.

Baldassano C, Beck DM, Fei-Fei L. 2015. Parcellating connectivity in spatial maps. PeerJ. doi:10.7717/peerj.784

Biederman I, Mezzanotte RJ, Rabinowitz JC. 1982. Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol. 14(2):143–177.

Borenstein E, Malik J. 2006. Shape guided object segmentation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p. 969–976.

Brass M, Schmitt R, Spengler S, Gergely G. 2007. Investigating action understanding: inferential processes verses action simulation. Curr Biol. 17:2117–2121.

Buccino G, Binkofski F, Fink G, Fadiga L, Fogassi L, Gallese V, Freund H-J. 2001. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. Eur J Neurosci. 13:400–404.

Buccino G, Binkofski F, Riggio L. 2004. The mirror neuron system and action recognition. Brain Lang. 89:370–376.

Buccino G, Lui F, Canessa N, Patteri I, Lagravinese G, Benuzzi F, Rizzolatti G. 2004. Neural circuits involved in the recognition of actions performed by nonconspecifics: an fMRI study. J Cogn Neurosci. 16(1):114–126.

Calvo-Merino B, Glaser D, Grèzes J, Passingham R, Haggard P. 2004. Action observation and acquired motor skills: an fMRI study with expert dancers. Cerebr Cortex. 15:1243–1249.

Calvo-Merino B, Grèzes J, Glaser D, Passingham R, Haggard P. 2006. Seeing or doing? Influence of visual and motor familiarity in action observation. Curr Biol. 16:1905–1910.

Caspers S, Zilles K, Laird A, Eickhoff S. 2010. ALE meta-analysis of action observation and imitation in the human brain. NeuroImage. 50:1148–1167.

Chang C, Lin C. 2011. LIBSVM: a library for support vector machines. ACM Trans Intelligent Systems Technol. 2(27):1–27.

Chong TT-J, Cunnington R, Williams M, Kanwisher N, Mattingley J. 2008. fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Curr Biol. 18:1576–1580.

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 29(3):162–173.

Csibra G. 2007. Action mirroring and action understanding: an alternative account. In: Haggard P, Rossetti Y, Kawato M, editors. Sensorimotor foundations of higher cognition. New York (NY): Oxford University Press, Inc. p. 435–459.

Culham JC, Valyear KF. 2006. Human parietal cortex in action. Curr Opin Neurobiol. 16:205–212.

Decety J, Lamm C. 2007. The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. Neuroscientist. 13:580–593.

DiCarlo J, Zoccolan D, Rust N. 2012. How does the brain solve visual object recognition? Neuron. 73(3):415–434.

Downing P, Peelen M. 2011. The role of occipitotemporal body-selective regions in person perception. Cogn Neurosci. 2: 186–226.

Eklund A, Nichols T, Knutsson H. 2015. Can parametric statistical methods be trusted for fMRI based group studies? arXiv:1511.01863.

Golarai G, Ghahremani DG, Whitfield-Gabrieli S, Reiss A, Eberhardt JL, Gabrieli JD, Grill-Spector K. 2007. Differential development of high-level visual cortex correlates with category-specific recognition memory. Nat Neurosci. 10:512–522.

Goldenberg G. 2009. Apraxia and the parietal lobes. Neuropsychologia. 47(6):1449–1459.

Green C, Hummel JE. 2006. Familiar interacting object pairs are perceptually grouped. J Exp Psychol Hum Percept Perform. 32:1107–1119.

Grézes J, Frith C, Passingham R. 2004. Brain mechanisms for inferring deceit in the actions of others. J Neurosci. 24:5500–5505.

Hauser M, Wood J. 2010. Evolving the capacity to understand actions, intentions, and goals. Annu Rev Psychol. 61:303–324.

Hein G, Knight R. 2008. Superior temporal sulcus - it's my area: or is it? J Cogn Neurosci. 20(12):2125–2136.

Hickok G. 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. J Cogn Neurosci. 21(7):1229–1243.

Huberle E, Karnath H-O. 2012. The role of temporo-parietal junction (TPJ) in global Gestalt perception. Brain Struct Funct. 217:735–746.

Kable JW, Chatterjee A. 2006. Specificity of action representations in the lateral occipitotemporal cortex. J Cogn Neurosci. 18:1498–1517.

Kaiser D, Strnad L, Seidl K, Kastner S, Peelen M. 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. J Neurophysiol. 111:82–90.

Kalénine S, Buxbaum LJ, Coslett HB. 2010. Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. Brain. 133:3269–3280.

Kim JG, Biederman I. 2011. Where do objects become scenes? Cereb Cortex. 21:1738–1746.

Kravitz D, Saleem K, Baker C, Ungerleider L, Mishkin M. 2013. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn Sci. 17(1):26–49.

Kubilius J, Baeck A, Wagemans J, Op de Beeck HP. 2015. Brain-decoding fMRI reveals how wholes relate to the sum of parts. Cortex. 72:5–14.

Lahnakoski JM, Glerean E, Salmi J, Jääskeläinen IP, Sams M, Hari R, Nummenmaa L. 2012. Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front Hum Neurosci. 6:233.

Lingnau A, Gesierich B, Caramazza A. 2009. Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. Proc Natl Acad Sci USA. 106(24):9925–9930.

MacEvoy SP, Epstein RA. 2011. Constructing scenes from objects in human occipitotemporal cortex. Nat Neurosci. 14:1323–1329.

MacEvoy SP, Epstein RA. 2009. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. Curr Biol. 19:943–947.

Marszalek M, Schmid C. 2007. Accurate object localization with shape masks. IEEE Conference on Computer Vision & Pattern Recognition.

Opelt A, Pinz A, Fussenegger M, Auer P. 2006. Generic object recognition with boosting. IEEE Trans Pattern Analysis Machine Intelligence. 28(3):416–431.

Pelphrey K, Morris J, McCarthy G. 2004. Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci. 16(10):1706–1716.

Riddoch MJ, Humphreys GW, Edwards S, Baker T, Willson K. 2003. Seeing the action: Neuropsychological evidence for action-based effects on object selection. Nat Neurosci. 6:82–89.

Rizzolatti G, Craighero L. 2004. The mirror-neuron system. Annu Rev Neurosci. 27:169–192.

Roberts K, Humphreys G. 2011. Action relations facilitate the identification of briefly-presented objects. Atten Percept Psychophys. 73:597–612.

Roberts KL, Humphreys GW. 2010. Action relationships concatenate representations of separate objects in the ventral visual system. NeuroImage. 52:1541–1548.

Saxe R. 2005. Against simulation: the argument from error. Trends Cogn Sci. 9(4):174–179.

Saxe R. 2006. Uniquely human social cognition. Curr Opin Neurobiol. 16(2):235–239.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". NeuroImage. 19:1825–1842.

Saxe R, Xiao D, Kovacs G, Perrett DI, Kanwisher N. 2004. A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia. 42:1435–1446.

Spunt R, Satpute A, Lieberman M. 2010. Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. J Cogn Neurosci. 23(1):63–74.

Stelzer J, Chen Y, Turner R. 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage. 65:69–82.

Tirilly P, Claveau V, Gros P. 2008. Language modeling for bag-of-visual words image categorization. New York, NY: International conference on Content-based image and video retrieval. p. 249–258.

Vul E, Harris C, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. Perspect Psychol Sci. 4(3):274–290.

Walther D, Caddigan E, Fei-Fei L, Beck D. 2009. Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci. 29(34):10573–10581.

Wang L, Mruczek REB, Arcaro MJ, Kastner S. 2014. Probabilistic maps of visual topography in human cortex. Cerebr Cortex. doi: 10.1093/cercor/bhu277.

Want SC, Harris PL. 2002. How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children. Dev Sci. 5(1):1–14.

Weiner KS, Grill-Spector K. 2011. Not one extrastriate body area: using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. NeuroImage. 56:2183–2199.

Wheatley T, Milleville SC, Martin A. 2007. Understanding animate agents: distinct roles for the social network and mirror system. Psychol Sci. 18(6):469–474.

Wiggett AJ, Downing PE. 2011. Representation of action in occipito-temporal cortex. J Cogn Neurosci. 23:1765–1780.

Xu J, Gannon PJ, Emmorey K, Smith JF, Braun AR. 2009. Symbolic gestures and spoken language are processed by a common neural system. Proc Natl Acad Sci USA. 106(49):20664–20669.

Yao B, Jiang X, Khosla A, Lin AL, Guibas LJ, Fei-Fei L. 2011. Action recognition by learning bases of action attributes and parts. Barcelona, Spain: International Conference on Computer Vision.

Yoon E, Humphreys G, Kumar S, Rotshtein P. 2012. The neural selection and integration of actions and objects: an fMRI study. J Cogn Neurosci. 24(11):2268–2279.

Zoccolan D, Cox DD, DiCarlo JJ. 2005. Multiple object response normalization in monkey inferotemporal cortex. J Neurosci. 25:8150–8164.