Behavioral/Cognitive

Neural Representations of Observed Actions Generalize across Static and Dynamic Visual Input

Alon Hafri, John C. Trueswell, and Russell A. Epstein
Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

People interact with entities in the environment in distinct and categorizable ways (e.g., *kicking* is *making contact with foot*). We can recognize these action categories across variations in actors, objects, and settings; moreover, we can recognize them from both dynamic and static visual input. However, the neural systems that support action recognition across these perceptual differences are unclear. Here, we used multivoxel pattern analysis of fMRI data to identify brain regions that support visual action categorization in a format-independent way. Human participants were scanned while viewing eight categories of interactions (e.g., *pulling*) depicted in two visual formats: (1) visually controlled videos of two interacting actors and (2) visually varied photographs selected from the internet involving different actors, objects, and settings. Action category was decodable across visual formats in bilateral inferior parietal, bilateral occipitotemporal, left premotor, and left middle frontal cortex. In most of these regions, the representational similarity of action categories was consistent across subjects and visual formats, a property that can contribute to a common understanding of actions among individuals. These results suggest that the identified brain regions support action category codes that are important for action recognition and action understanding.

Key words: **action recognition; fMRI; inferior parietal; multivoxel pattern analysis; occipitotemporal; premotor**

Significance Statement

Humans tend to interpret the observed actions of others in terms of categories that are invariant to incidental features: whether a girl pushes a boy or a button and whether we see it in real-time or in a single snapshot, it is still *pushing*. Here, we investigated the brain systems that facilitate the visual recognition of these action categories across such differences. Using fMRI, we identified several areas of parietal, occipitotemporal, and frontal cortex that exhibit action category codes that are similar across viewing of dynamic videos and still photographs. Our results provide strong evidence for the involvement of these brain regions in recognizing the way that people interact physically with objects and other people.

Introduction

The ability to recognize actions performed by others is crucial for guiding intelligent behavior. To perceive categories of actions, one must have representations that distinguish between them (e.g., *biting* is different from *pushing*) yet show invariance to different instantiations of the same action. Although previous work has described a network of regions involved in coding observed actions (the "action observation network" or AON; Caspers et al., 2010; Rizzolatti and Sinigaglia, 2010; Kilner, 2011; Urgesi et al., 2014), the extent to which these regions abstract across differences between action exemplars is not well understood.

Previous research has addressed the question of abstraction (i.e., invariance) in two ways. First, many neuroimaging and neuropsychological studies have explored generalization between observed and executed actions in an effort to resolve a debate over motor system involvement in action understanding (Chong et al., 2008; Dinstein et al., 2008; Kilner et al., 2009; Oosterhof et al., 2012a, 2012b; Tarhan et al., 2015; Tucciarelli et al., 2015; for review, see Rizzolatti and Sinigaglia, 2010; Oosterhof et al., 2013; Caramazza et al., 2014). Second, other studies have examined invariance to different perceptual instantiations of observed actions (Kable and Chatterjee, 2006; Oosterhof et al., 2012a; Watson et al., 2014; Tucciarelli et al., 2015). In an especially direct test of such invariance, Wurm and Lingnau (2015) found that representations in several AON regions distinguished between *opening*

Received Aug. 5, 2016; revised Jan. 16, 2017; accepted Feb. 3, 2017.

Author contributions: A.H., J.C.T., and R.A.E. designed research; A.H. performed research; A.H. analyzed data; A.H., J.C.T., and R.A.E. wrote the paper.

This work was supported by the Center for Functional Neuroimaging at the University of Pennsylvania; the National Science Foundation (NSF Integrative Graduate Education and Research Traineeship and NSF Graduate Research Fellowship to A.H.); and the National Institutes of Health (Vision Training Grant 2T32EY007035-36 to A.H.). We thank Rachel Olvera and Jennifer Deng for assistance with stimulus collection; Jack Ryan and Stamati Liapis for assistance with data collection; Michael Bonner, Steven Marchette, and Anjan Chatterjee for helpful comments on an earlier version of this manuscript; and the two actors who appeared in the videos.

The authors declare no competing financial interests.

Correspondence should be addressed to Alon Hafri, Department of Psychology, University of Pennsylvania, 425 S. University Avenue, Philadelphia, PA 19104. E-mail: ahafri@sas.upenn.edu.

DOI:10.1523/JNEUROSCI.2496-16.2017

 0270-6474/17/373056-16$15.00/0

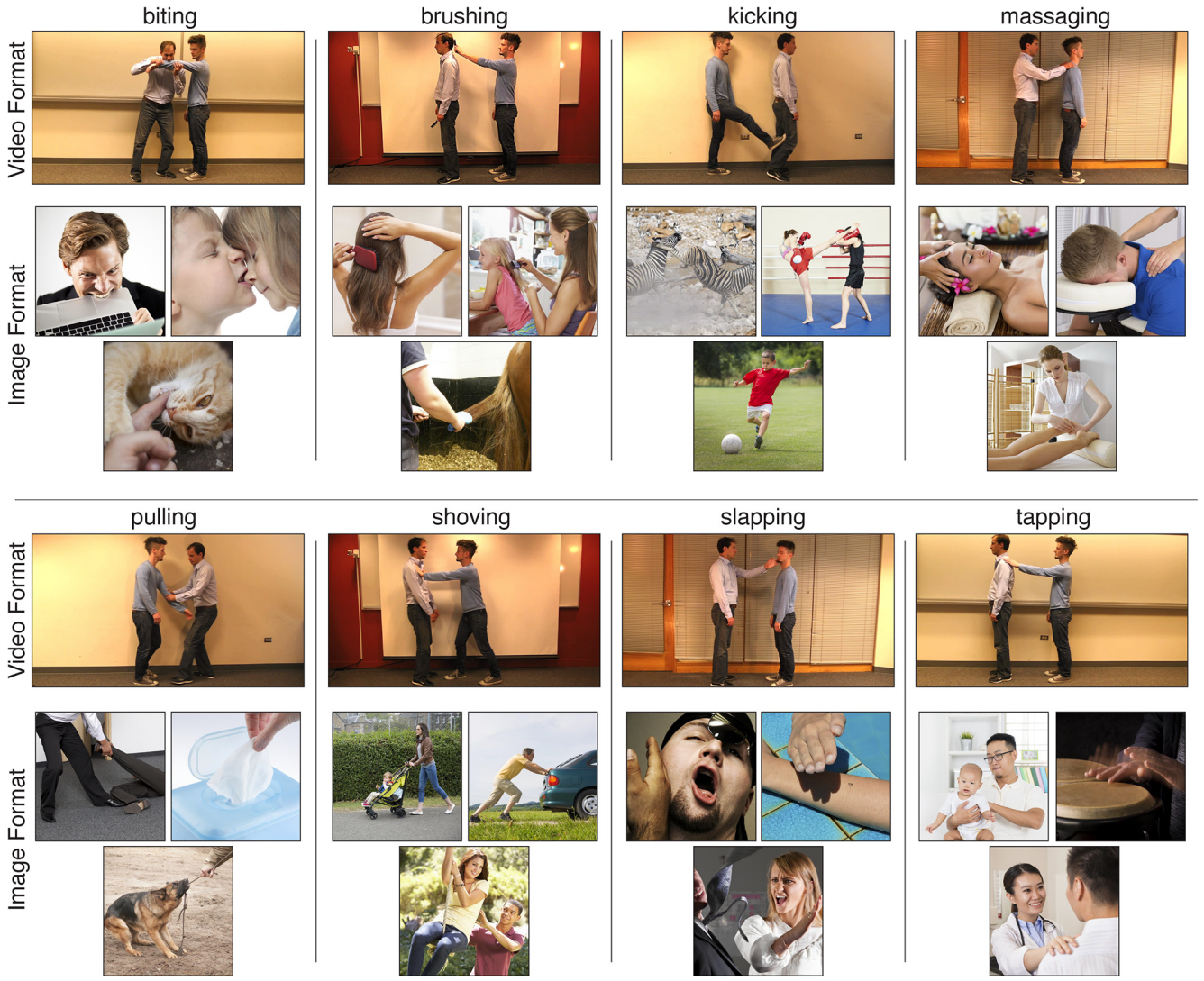

Figure 1. Examples of stimuli. Subjects viewed dynamic videos and still images of eight categories of interactions. For each action category, one still frame for the video format and three photographs for the image format are shown. In the video format, actor role (Agent/Patient), action direction (left/right), and scene background (four indoor backgrounds) were fully crossed within each action category. For example, in the *brushing* still frame depicted here, the blue-shirted actor is the Agent, the action direction is toward the left, and the background is the red wall, whereas in other *brushing* videos, this combination of factors was different (e.g., action direction toward the right instead of left). In the still image format, photographs from the internet were chosen to maximize variation in actors, objects, viewpoint, and scene context within each category. Image format examples shown here are photographs that we have license to publish and closely resemble the actual stimuli used.

and *closing* in a manner that generalized across different kinematic manipulations and acted-upon objects (i.e., across bottles and boxes; cf. Wurm et al., 2016). These findings and others suggest that at least a subset of AON regions support abstract codes for actions that could conceivably facilitate perceptual recognition.

However, we posited that an action recognition system should display two additional kinds of perceptual generalization. First, it should support representations of action category that are invariant not only to the acted-on object or kinematics, but also to other incidental perceptual features such as the identities of entities involved and location. Whether a girl pushes a boy or a boy pushes a button, and whether it takes place in a classroom or a playground, it is still *pushing*. Second, these representations should be elicited both by dynamic visual sequences, in which the entire action is observed, and static snapshots, from which the causal sequence must be inferred (Hafri et al., 2013). Several studies have found action-specific representations using static images (Ogawa and Inui, 2011; Watson et al., 2014) but, crucially, none has demonstrated common representations across dynamic and static input. Beyond testing these invariances, we also wished to examine actions performed with a wide variety of effectors (e.g., foot, mouth), not just hand/arm actions that are commonly investigated in the literature (see also Kable and Chatterjee, 2006; Jastorff et al., 2010; Watson et al., 2014).

To these ends, we used multivoxel pattern analysis (MVPA) of fMRI data to identify regions supporting abstract action representations. We scanned subjects while they viewed eight action categories in two visual formats (Fig. 1): (1) visually controlled videos of two interacting actors and (2) photographs involving a variety of actors, objects, and settings. We then attempted to decode action category by comparing multivoxel patterns across the formats, which should be possible in regions that support action category representations not tied to low-level visual features correlated with actions. To anticipate, we were able to decode action category across visual formats in bilateral inferior

parietal lobule (IPL), bilateral occipitotemporal cortex (OTC), left premotor cortex, and left middle frontal gyrus (mFG). We then conducted further analyses in these regions to probe the stability of their representations across perceptual features and subjects. Finally, we tested for action decoding in independently localized functional OTC regions to determine their involvement in action representation (Kanwisher, 2010). Together, our results support the hypothesis that AON regions contain neural populations that can mediate action recognition regardless of the dynamicity of visual input and the perceptual details of the observed action.

Materials and Methods

Participants

Fifteen healthy adults (8 female; mean age 22.1 ± 4.6 years; range 18–35 years) were recruited from the University of Pennsylvania community. All participants were healthy, had normal or corrected-to-normal vision, and provided written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board. All were right-handed, except one who was ambidextrous. All were native English speakers and one was bilingual. Data from an additional participant was discarded before analysis for an inability to complete the entire experiment.

For selection of video stimuli, a group of 16 additional individuals (University of Pennsylvania undergraduates) participated in an online norming survey for psychology course credit. For selection of still image stimuli, 647 individuals on Amazon's Mechanical Turk (MTurk) participated in a separate online norming survey. All MTurk workers were located in the United States and had 90–95% worker approval rate on previous tasks.

For the eye-tracking control experiment, a group of 16 additional individuals (University of Pennsylvania undergraduates) participated for psychology course credit.

Stimuli

To identify neural representations of action categories that were invariant to incidental perceptual features, we scanned subjects while they viewed eight different categories of interactions: *biting*, *brushing*, *kicking*, *massaging*, *pulling*, *shoving*, *slapping*, and *tapping*.

The action categories were viewed in two formats: visually controlled video clips created in the laboratory and visually varied photographic images taken from the internet. The use of videos allowed us to examine action representations elicited by dynamic stimuli, thus mimicking action perception in the natural world. This approach is the standard in previous literature investigating action recognition (Grossman and Blake, 2002; Vangeneugden et al., 2014; Wurm and Lingnau, 2015). The use of images allowed us to determine whether the same action representations were elicited even when actions are perceived from a static snapshot, which has been shown in previous behavioral studies to be sufficient for recognition even from brief displays (Hafri et al., 2013).

In addition, by using one format that was more visually controlled (the videos) and another that was more visually varied (the images), we decreased the possibility of potential confounding factors present in either format alone. The videos always contained the same set of actors and scene contexts, so the different body movement patterns were the only aspect of the stimuli that allowed categories to be discriminated (apart from *brushing*, which contained a unique object). Although this had the merit that distinctions between categories within the videos could not be attributed to differences in actors or scene context, it had the disadvantage that category was inevitably confounded with lower-level motion properties that covaried with the actions. Conversely, in the still images, distinctions between categories could not be attributed to low-level motion patterns; however, because the stimuli were less visually constrained, it remained possible that action category could have covaried with the presence of particular types of actors, objects, scene contexts, or even implied motion (Kourtzi and Kanwisher, 2000; Senior et al., 2000). By comparing patterns of fMRI responses to the videos with those to the still images when identifying category representations, we reduced these concerns because the most likely confounds in one stimulus set are either absent or controlled for in the other.

Video stimuli. A total of 128 video clips (2.5 s each) were filmed and divided equally into eight action categories. A pair of male actors of similar height performed all interactions. Video clips were filmed in front of four different indoor backgrounds; one actor appeared as the Agent (i.e., the entity that performs an action on another entity) and the other as the Patient (i.e., the entity on which an action is performed) and the action was directed either toward the left or to the right. These three factors were crossed to make 16 video clips for each category: four backgrounds × two actor roles (actor A as Agent or actor B as Agent) × two action directions (leftward or rightward). For example, for *biting*, there were four video clips (with different backgrounds) of actor A on the left biting actor B on the right, four of A on the right biting B on the left, four of B on the left biting A on the right, and four of B on the right biting A on the left.

The two actors were centered in the video frame in full-body profile view and started each clip at rest with arms at their sides. For half of the action categories (*biting*, *pulling*, *shoving*, and *slapping*), the actors faced one another and, for the other half (*brushing*, *kicking*, *massaging*, and *tapping*), they both faced the same direction. For *brushing*, both actors always held a brush. Actors kept neutral faces throughout the duration of the videos. Example still frames for each action category appear in Figure 1.

To ensure that our videos could be easily interpreted as depicting the intended action categories, we obtained descriptions of our videos from a separate group of raters. These participants viewed a random selection of 100 videos one at a time and provided a verbal label that in their opinion best described each action depicted (total 15 labels per video clip, SD = 0.45, range 14–16). These verbal labels confirmed that our video clips depicted the intended action categories: all were described with the intended verbal label or close synonym >95% of the time. Synonyms included: for *biting*: *chomping*, *gnawing*; for *brushing*: *combing*; for *kicking*: none; for *massaging*: *rubbing*; for *pulling*: *yanking*, *tugging*, *grabbing*, *dragging*; for *shoving*: *pushing*; for *slapping*: *hitting*, *smacking*; and for *tapping*: *patting*.

Still image stimuli. For each action category, we used 16 still images (128 total), which were selected to maximize the within-category variety of actors, objects, and scene contexts (e.g., only one *biting* image included a person biting an apple). Stimuli included both animate and inanimate Patients (the entity on which an action is performed).

To create this stimulus set, an initial set of candidate stimuli were obtained from Google Images using search terms that included the target verbal label, close synonyms, and short phrases (e.g., *patting* or *patting on the back* for *tapping*, *combing* for *brushing*, *pushing* for *shoving*, and *smacking in the face* for *slapping*). This search procedure yielded 809 images (87–118 images per category). To reduce this set, a group of MTurkers followed the same norming procedure as for the videos. Each viewed a random selection of 60 images and provided a verbal label that best described each action depicted (total 16 labels per image, SD = 1.6, range 11–20). Based on these labels, we eliminated images that did not have high name agreement with the target verbal label or close synonym. Synonyms included: for *biting*: *gnawing*, *tasting*, *eating*; for *brushing*: *combing*; for *kicking*: *kickboxing*; for *massaging*: *rubbing*, *back-rubbing*; for *pulling*: *yanking*, *tugging*, *grabbing*, *grasping*, *dragging*; for *shoving*: *pushing*; for *slapping*: *hitting*, *smacking*, *punching*; and for *tapping*: *patting*, *poking*, *touching*. Name agreement was at least 87% for each *biting*, *brushing*, *kicking*, and *massaging* image. For the other categories (*pulling*, *shoving*, *slapping*, and *tapping*), the name agreement criterion was relaxed to a minimum of 75%, 75%, 64%, and 53%, respectively, to retain at least 16 images per category. This resulted in a set of 209 images (16–38 per category) with high name agreement.

We then calculated three measures to assess low-level visual similarity among the remaining images with the aim of choosing a final image set with maximal visual dissimilarity within each category. The first measure was the Gist model (Oliva and Torralba, 2001), which is a set of image descriptors that represent the energy at different spatial frequencies and scales. Image similarity was calculated as the correlation of descriptor magnitudes between each pair of images. The other two measures were

the average hue, saturation, value (HSV) hue channel values for each image and the average HSV saturation channel values for each image. With these three measures in hand, we ran 10,000 permutations in which we randomly selected a subset of 16 images per category and calculated, for each category, the average distance in Gist space between all 16 images and the variance across images in the hue and saturation channels. Of these permutations, we selected the one with the greatest average within-category Gist distance and greatest within-category variance across images for hue and saturation. Across the final set of 128 images, we luminance matched the HSV value channel using the MATLAB SHINE toolbox (Willenbockel et al., 2010) and converted the images back to RGB space. Examples for each action category appear in Figure 1.

MRI acquisition

Scanning was performed at the Center for Functional Imaging at the University of Pennsylvania on a 3T Siemens Prisma scanner equipped with a 64-channel head coil. High-resolution T1-weighted images for anatomical localization were acquired using a 3D magnetization-prepared rapid acquisition gradient echo pulse sequence [repetition time (TR), 1620 ms; echo time (TE), 3.09 ms; inversion time, 950 ms; voxel size, $1 \times 1 \times 1$ mm; matrix size, $192 \times 256 \times 160$ mm]. T2*-weighted images sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a gradient echo echoplanar pulse sequence (TR, 3000 ms; TE, 30 ms; flip angle, 90°; voxel size, $3 \times 3 \times 3$ mm; field of view, 192 mm; matrix size, $64 \times 64 \times 44$). Visual stimuli were displayed at the rear bore face on an InVivo SensaVue Flat Panel Screen at 1920×1080 pixel resolution (diagonal = 80.0 cm, width × height = 69.7×39.2 cm). Participants viewed the stimuli through a mirror attached to the head coil. Images subtended a visual angle of $\sim 11.7 \times 11.7°$ and videos subtended a visual angle of $\sim 18.9 \times 10.7°$. Responses were collected using a fiber-optic button box.

Design and task

Main experiment. To determine BOLD response to action categories in different visual formats, participants were scanned with fMRI while viewing the still images and videos. Images and videos were presented in separate scan runs, with four runs per format (eight total), alternating in sets of two (e.g., image run 1, image run 2, video run 1, video run 2, image run 3, etc.). The format that appeared first was counterbalanced across participants. Within format, stimuli were identical within odd-numbered and within even-numbered runs (e.g., stimuli in video runs 1 and 3 were identical, stimuli in image runs 2 and 4 were identical, etc.). Therefore, except for repetition trials (see next paragraph), each stimulus was shown a total of two times over the course of the experiment (in separate runs).

To ensure attention to the stimuli, participants were instructed to press a button whenever the stimulus on the current trial was exactly the same as the stimulus on the immediately preceding trial (repetition trials). Importantly, this task could not be performed by attention to the action category alone. Trials occurred every 3 s in a rapid event-related design. Videos were displayed for 2500 ms, followed by a 500 ms intertrial interval (ITI) with a white fixation cross centered on a gray background. Images were displayed for 1200 ms, followed by an 1800 ms ITI. Each scan run included 64 trials in which unique stimuli were shown (8 for each category), 8 repetition trials, and 12 null trials, in which participants viewed a blank screen with a fixation crosshair for 3 s (total duration 4 min 33 s per scan run). A unique pseudorandomized sequence of stimuli was generated for each scan run using *optseq2* ((http://surfer.nmr.mgh.harvard.edu/optseq; RRID:SCR_014363) with the following parameters: psdwin 0 to 21, nkeep 10000, focb 100, nsearch 200000. Five extra null trials were added at the end of each scan run to ensure that we captured the hemodynamic response to the last stimulus in each run.

Video stimuli were divided such that odd video runs contained the videos with two of the four backgrounds and even video runs contained the videos with the remaining two backgrounds. Therefore, each video run included two stimuli for each combination of action category, actor roles, and action direction (eight stimuli per action category in each video run). The combinations of background splits were cycled through for each subject (e.g., subject 1 had backgrounds 1 and 2 in odd runs and backgrounds 3 and 4 in even runs, subject 2 had backgrounds 1 and 3 in odd runs and backgrounds 2 and 4 in even runs, etc.). Image stimuli were assigned to odd and even runs with a unique split for each subject (eight images per category for the odd runs and eight per category for the even runs). Stimuli were displayed using a Macbook Pro laptop with MATLAB version 2013b (The MathWorks; RRID:SCR_001622) and the MATLAB Psychophysics Toolbox version 3.0.11 (Brainard, 1997; Pelli, 1997; RRID:SCR_002881).

Functional localizers. To determine the information content for action categories in functionally selective brain regions, all subjects completed three functional localizer scans in the middle of each scan session. The first localizer featured static image stimuli to identify regions responsive to different stimulus categories. This run consisted of 25 blocks (15 s long each; run duration 6 min 15 s) of static images of faces, objects, scrambled objects, bodies, and scenes. Blocks 1, 7, 13, 19, and 25 were null blocks with a blank gray screen and white crosshair. Images were presented for 800 ms, each followed by a 200 ms interstimulus interval. Subjects performed a one-back repetition detection task (two repetitions per block).

The second localizer featured dynamic stimuli to identify regions responsive to biological motion and basic motion (Grossman et al., 2000; Vaina et al., 2001; Grossman and Blake, 2002). This run consisted of 25 blocks (18 s long each; run duration 7 min 30 s) of intact point-light displays of single-person actions (e.g., *waving*, *jumping*), scrambled versions of these stimuli (in which motion patterns were preserved but starting position of points was randomized), and static point-light still frames randomly selected from the scrambled point-light videos. Blocks 1, 5, 9, 13, 17, 21, and 25 were null blocks with a blank gray screen and centered red fixation point. Stimuli were presented for 1500 ms each, with a 300 ms interstimulus interval. Subjects performed a one-back repetition detection task (one repetition per block). To create these stimuli, motion capture data were taken from the Carnegie Mellon Motion Capture Database (http://mocap.cs.cmu.edu) and animated using the Biomotion toolbox (van Boxtel and Lu, 2013).

The third localizer featured linguistic stimuli to identify regions responsive to linguistic depictions of actions (design based on Bedny et al., 2008). This run consisted of 20 blocks (18 s long each; run duration 6 min 36 s) in which verbs and nouns were presented visually to participants in separate alternating blocks. On each trial (2.5 s each), participants had to rate the similarity in meaning of 2 words presented sequentially (1 s each) by performing a button press indicating their response on a scale of 1 to 4. Words were a set of 50 motion verbs (e.g., *to stumble*, *to prance*) and 50 animal nouns (e.g., *the gorilla*, *the falcon*) approximately equated for similarity and difficulty (available in supplementary material in Bedny et al., 2014). Words were paired randomly within block.

fMRI data analysis

Overview. Our primary goal was to identify representations of action categories that generalized across dynamic videos and static images. To identify brain regions supporting such representations, we implemented a whole-brain searchlight analysis of multivoxel responses to action categories shown in both video and image format. Once these regions were identified, we performed several further analyses to determine the properties of the encoded action categories. First, we compared the cross-format searchlight results to results from within-format searchlight analyses to observe the degree of overlap of within- and cross-format decoding. Second, with the regions identified by the cross-format searchlights, we performed a more fine-grained analysis of the responses to the video stimuli to test whether category representation elicited by videos generalized across actor role and direction of action. Third, we performed a representational similarity analysis within these regions to determine whether their category spaces were similar across subjects. Finally, to determine the relationship between functional selectivity and coding of action category, we tested for cross-format and within-format category decoding in a number of functional regions of interest (ROIs) defined based on univariate responses in localizer scans.

Data preprocessing. Functional images were corrected for differences in slice timing by resampling slices in time to match the first slice of each volume. Images were then realigned to the first volume of the scan and subsequent analyses were performed in the subject's own space. Motion

correction was performed using MCFLIRT (Jenkinson et al., 2002). Data from the functional localizer scans were smoothed with a 5 mm full-width at half-maximum Gaussian filter; data from the main experimental runs were not smoothed.

Whole-brain analysis of cross- and within-format action category decoding. To search for cross-format action category information across the brain, we implemented a searchlight analysis (Kriegeskorte et al., 2006) of multivoxel patterns elicited by the eight action categories in video and static image format. We centered a small spherical ROI (radius 5 mm, 19 voxels) around every voxel of the brain separately for each participant and then calculated a discrimination index within each sphere. This index was defined as the difference between the Pearson correlation across scan runs for patterns corresponding to the same action category in different formats (e.g., *kicking* in the video format with *kicking* in the image format) and the Pearson correlation across scan runs for patterns corresponding to different action categories in different formats (e.g., *kicking* in the video format with *brushing* in the image format). If this index is positive, then this indicates that the searchlight sphere contains information about action category (Haxby et al., 2001). We then assigned the resulting value to the central voxel of the sphere.

To define the activity patterns, we used general linear models (GLMs) implemented in FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki; RRID:SCR_002823) to estimate the response of each voxel to each action category in each scan run. Each runwise GLM included one regressor for each action category (eight total), one regressor for repetition trials, regressors for six motion parameters, and nuisance regressors to exclude outlier volumes discovered using the Artifact Detection Toolbox (http://www.nitrc.org/projects/artifact_detect; RRID:SCR_005994). A high-pass filter (100 Hz) was used to remove low temporal frequencies before fitting the GLM and the first two volumes of each run (always extra null trials) were discarded to ensure data quality. Individual patterns for each run were normalized before cross-run comparison by calculating the *z*-score for each voxel, across conditions. *Z*-scored patterns were averaged within odd and within even runs of the same format (e.g., image runs 1 and 3 were averaged; video runs 2 and 4 were averaged) and discrimination index scores were calculated based on correlations between even and odd sets of runs.

To produce optimal alignment of searchlight maps across subjects, we first reconstructed anatomical pial surface and gray-white matter boundaries for each subject using FreeSurfer version 5.3.0 (http://surfer.nmr.mgh.harvard.edu; RRID:SCR_001847). These were aligned to a FreeSurfer standard template using a spherical transformation (Fischl et al., 1999) and, based on this alignment, the *mri_vol2vol* tool was used to calculate registrations from subject functional space to FreeSurfer standard. These standard-space subject maps were submitted to a second-level random-effects analysis in FSL. To correct for multiple comparisons, the group-level *t*-map was first submitted to threshold-free cluster enhancement (TFCE; Smith and Nichols, 2009), an algorithm designed to offer the sensitivity benefits of cluster-based thresholding without the need for an arbitrarily chosen threshold. The TFCE statistic represents the cluster-like local support for each voxel using empirically and theoretically derived height and extent parameters. This TFCE map was then whole-brain corrected ($p < 0.05$) for the familywise error rate using standard permutation tests implemented in FSL with the *randomize* function (10,000 permutations) and spatial 5 mm FWHM variance smoothing, which is recommended for $df < 20$ because it reduces noise from poorly estimated SDs in the permutation test procedure (Nichols and Holmes, 2002).

Searchlight analyses were also conducted within visual format (one for Image Format, one for Video Format). The same analyses as above were implemented, except for the following. For Image Format, patterns were compared between image runs only (e.g., *kicking* in the odd image runs with *kicking* in the even video runs); for Video Format, between video runs only (e.g., *kicking* in the odd video runs with *kicking* in the even image runs). To compare the overlap of within- and cross-format decoding regions qualitatively, we overlaid whole-brain searchlight maps for the different format comparisons to examine regions of conjunction. Here, the maximum *p*-value (TFCE, whole-brain corrected) is the valid value for conjunction inference in each voxel (the minimum statistic compared with the conjunction null; Nichols et al., 2005).

Cross-format ROI definition. We used the results of the cross-format searchlight analysis to define ROIs for two subsequent analyses, described below. ROIs were constructed by taking the intersection of the cross-format decoding map (whole-brain corrected) and spheres centered on the cluster peaks from this map (Fairhall and Caramazza, 2013) and transforming the defined region back into the native functional space for each subject. Because spheres with a given radius may yield different ROI sizes after intersection with the whole-brain map, the radius of these spheres was adjusted separately for each region so that ~100 voxels were contained within each ROI after transformation to subject space (mean 108 voxels, SD = 15, range 81–156).

Invariance to controlled factors in the video stimuli. The first follow-up analysis tested whether the patterns elicited by the videos showed invariance to incidental properties of the actions, such as the action direction (leftward vs rightward) and actor roles (actor A as Agent or actor B as Agent). To test whether this was the case, we implemented additional GLMs that included one regressor for each action category × action direction × actor role combination within each video run (32 regressors total per run, with two video stimuli contributing to each estimate). Multivoxel patterns within run were *z*-scored across the 32 conditions and averaged within odd and within even runs. For each cross-format ROI, pairwise Pearson correlations were calculated for patterns between all 32 conditions across odd and even runs and correlation coefficients were averaged for all combinations of same versus different action category, same versus different action direction, and same versus different actor roles, yielding eight mean correlations values per subject and ROI. These pattern similarity values were then entered into 2 × 2 × 2 repeated-measures ANOVAs (one for each ROI), with action category, action direction, and actor roles as factors. Early visual cortex (EVC, defined in a separate functional localizer described below) was also included in this analysis for comparison with the cross-format ROIs. *P*-values for *F* statistics were corrected for multiple comparisons across the nine ROIs using the Bonferroni–Holm method separately for each set of *F* statistics yielded by the ANOVA. The Bonferroni–Holm method is uniformly more powerful than standard Bonferroni while still controlling for the familywise error rate (Holm, 1979). Note that, although the same Video Format data were used for cross-format ROI definition and for this follow-up analysis, this analysis is unlikely to be affected by circular analysis problems (Kriegeskorte et al., 2009) because the cross-format ROI definition procedure used GLMs that collapsed across actor role and action direction for each action category. Therefore, the Video patterns used in the cross-format ROI definition procedure did not contain information about actor role or action direction.

Representational similarity analysis. The second follow-up analysis tested whether the patterns that allow for action discrimination within individual reflect a common representational space across individuals; that is, whether actions are represented in a similar way from person to person. To examine this issue, we used representational similarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013). Within each cross-format ROI, representational dissimilarity matrices (RDMs) were constructed using the pairwise Pearson correlation distances ($1 - r$) between multivoxel patterns for each action category to every other. Three separate RDMs were constructed for every subject and ROI: a video format RDM (even to odd video run correlations), an image format RDM (even to odd image run correlations), and a cross-format RDM (all video to all image run correlations). Cross-subject consistency in representational space was then assesses by calculating the Spearman correlation between each subject's RDM and every other subject's RDM separately for video, image, and cross-format RDMs. Because we were interested in similarities and differences between categories, rather than reliability within categories, only off-diagonal elements of the RDMs were included in the calculation. These intersubject correlations represent the similarity in representational space from each subject to every other, abstracted away from the underlying units of representation (voxels). If the mean intersubject correlation is significantly above zero, it indicates that the relationship among representational spaces is reliable.

Because the intersubject RDM correlation values were not independent (i.e., RDMs from each subject were used more than once in the intersubject RDM comparisons, e.g., subject 1 to subject 2, subject 1 to subject 3, etc.), permutation tests were used to determine chance levels. In these tests, for each comparison type (video, image, cross), the condition labels of the subject RDMs were shuffled before calculating the intersubject correlations. The mean intersubject correlation was calculated across all pairwise subject comparisons 10,000 times for each comparison type and cross-format ROI. The *p*-value was simply the proportion of times the true mean intersubject correlation was lower than a permuted intersubject correlation. These *p*-values were then corrected for multiple comparisons across the eight ROIs using the Bonferroni–Holm method separately for each comparison type. The mean chance intersubject correlation from permutation testing was approximately zero in all cases (mean 7.77×10^{-5}, range -1.50×10^{-3} to 1.50×10^{-3}, across all ROIs and comparison types).

Note that although the same data were used for cross-format ROI definition and for this follow-up analysis, the results of these analyses do not follow trivially from the finding of cross-format action category representations in these regions. In particular, because the action category discrimination index was quantified separately for each subject (using each subject's own representational space), reliable action category decoding across subjects does not logically ensure that their representational spaces will be related to one another. To confirm this point, we ran a simulation using randomly constructed RDMs. We observed no correlation between the magnitude of the discrimination indices and the Spearman correlation of the off-diagonal values across RDMs (mean -6.48×10^{-4}, SD = 0.03, across 1000 simulations).

Functionally localized ROIs. We also examined action decoding in several functional ROIs that previous work suggested might play a role in processing actions or perceptual constituents of actions. These ROIs were defined based on fMRI responses during three functional localizer scans (described above).

Data from the first localizer scan were used to define ROIs related to the viewing of specific stimulus categories, using a group-constrained subject-specific (GSS) ROI definition method (Julian et al., 2012). This approach yields similar individual subject functional ROIs to the traditional hand-drawn ROI pipeline, but uses an objective and automatic method. Each ROI was initially defined in each subject as the top 100 voxels in each hemisphere that responded more to the contrast of interest and fell within the group-parcel mask for the given ROI. Parcel masks were derived from a large number of separate subjects undergoing similar localizers using this method (parcels are available at http://web.mit.edu/bcs/nklab/GSS.shtml). Using this method, we identified the following ROIs, each using the contrast listed in parentheses: EVC (scrambled objects > objects); lateral occipital (LO) and posterior fusiform (pFs; objects > scrambled objects); occipital face area (OFA), anterior fusiform face area (FFA), and right posterior FFA (faces > objects); extrastriate body area (EBA) and right fusiform body area (FBA; bodies > objects); and occipital place area (OPA), parahippocampal place area (PPA), and retrosplenial complex (RSC; scenes > objects).

Data from the second localizer scan (dynamic stimuli) were used to define two motion-sensitive functional ROIs. GSS parcels were not available for these stimulus contrasts, so these ROIs were hand drawn. Human middle temporal complex (hMT+) was defined as the set of contiguous voxels responding more to scrambled than static point-light displays in the vicinity of the posterior inferior temporal sulcus separately in both hemispheres. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t >$ 5.3, range 3–8). The biological-motion-selective posterior superior temporal sulcus (pSTS-bio) was defined as the set of contiguous voxels responding more to intact than scrambled point-light displays in the vicinity of the posterior superior temporal sulcus in the right hemisphere. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t > 2.9$, range 2.0–4.7, identified in 11 of 15 participants).

Data from the third localizer scan (linguistic stimuli) were used to define the verb-selective left posterior middle temporal gyrus (pMTG-verb), defined as the set of contiguous voxels responding more to verbs than nouns in the vicinity of the left posterior middle temporal gyrus. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t > 3.7$, range 2.4–4.5, identified in 11 of 15 participants).

Because these functional ROIs often partially overlapped in individual subjects, we excluded voxels falling into more than one ROI (cf. Schwarzlose et al., 2008; Weiner and Grill-Spector, 2013). This allowed us to isolate the specific contribution of voxels with certain functional profiles (e.g., body-selective or motion-selective), without contamination from nearby regions with different functional profiles. After these exclusions, the mean size of the ROIs was as follows: EVC: 186 voxels (SD = 15, range 150–200); hMT+: 146 voxels (SD = 30, range 98–220); pSTS-bio: 51 voxels (SD = 22, range 16–93); LO: 155 voxels (SD = 15, range 134–172); pFs: 142 voxels (SD = 21, range 86–163); anterior FFA: 150 voxels (SD = 17, range 122–178); right posterior FFA: 200 voxels (no overlap); OFA: 165 voxels (SD = 23, range 114–193); EBA: 116 voxels (SD = 24, range 87–160); right FBA: 65 voxels (SD = 13, range 43–92); OPA: 195 voxels (SD = 3, range 190–200); PPA: 181 voxels (SD = 14, range 147–197); RSC: 200 voxels (no overlap); and pMTG-verb: 94 voxels (SD = 60, range 35–211). Analyses using ROIs in which overlapping voxels were not excluded yielded qualitatively similar results.

Action category discrimination indices for the video, image, and cross-format comparisons were calculated separately within each ROI for each subject and submitted to two-tailed one-sample *t* tests against chance (zero). *P*-values were corrected for multiple comparisons across functional ROIs separately within comparison type using the Bonferroni–Holm method (14 tests for each comparison type).

Eye-tracking control task

To ensure that action category decoding could not be attributed to differences in spatial attentional allocation, we ran a control study in which a separate group of participants underwent the identical procedure as in the main fMRI experiment, but outside of the scanner and while their gaze location was recorded by a remote binocular eye tracker situated within the visual display monitor (Tobii T120 eye tracker sampling at 60 Hz).

2D gaze maps were created for each combination of subject, format (Image or Video), run (four per format), and action category (eight) by binning gaze locations on the screen into 70 horizontal × 56 vertical bins. In other words, gaze maps akin to a 2D histogram were formed by dividing the screen extent into 70 × 56 bins and each eye-tracking sample was placed into its corresponding location in this set of bins (ignoring the time dimension). As with the fMRI voxel patterns, these gaze maps were *z*-scored across action category (for each subject, format, and run) and even and odd run maps were averaged together. We then attempted to decode action category both within- and across-format using the 2D gaze maps. Pearson correlations were calculated between even- and odd-run gaze maps corresponding to each action category (for each subject and analysis type separately). The discrimination index was the average within-category correlation minus the average between-category correlation. We tested the significance of this discrimination index across subjects separately for Image Format, Video Format, and Cross Format.

Results

Behavioral performance

One participant reported that she misunderstood the instructions for her first video run, so data for this run (behavioral and imaging) were excluded. For the remaining data, behavioral performance on the one-back repetition detection task was good, indicating that participants were playing close attention to the stimuli. For image runs, the mean accuracy on repetition trials was 0.91 (SD = 0.08), the mean false alarm rate was 0.002 (SD = 0.002), and the average reaction time (RT) on correct trials was 694 ms (SD = 82 ms). For video runs, the mean accuracy was 0.89 (SD = 0.10), the mean false alarm rate was 0.014 (SD = 0.015), and the average RT on correct trials was 1117 ms (SD = 157 ms).

Table 1. MNI locations, extent, mean cross-format discrimination index, significance, and peak statistics for the clusters identified in the cross-format action category searchlight, ordered by cluster extent (number of voxels)

Cluster						Peak					
Region	Extent	Cross-format discrimination index	*x*, *y*, *z* (center of gravity)			$p_{(FWE\text{-}cor)}$	Pseudo-*t*	$p_{(unc)}$	*x*, *y*, *z*		
Left IPL (ventral)[a]	605	0.045	−49.2	−26.7	38.7	0.002	5.07	1E-04	−58	−37	28
Left premotor (ventral)[b]						0.003	3.85	9E-04	−55	−4	40
Left IPL (dorsal)						0.02	5.73	3E-04	−40	−43	49
Left postcentral						0.003	5.03	3E-04	−58	−22	37
Left premotor (dorsal)						0.031	3.56	5E-04	−31	−4	43
Right LOTC[a]	96	0.049	48.9	−61.5	6.1	0.004	6.11	2E-04	44	−61	4
Right IPL[a]	64	0.045	56.3	−27.9	37.3	0.016	4.91	3E-04	53	−28	37
Left LOTC[a]	28	0.050	−42.9	−80.6	−0.1	0.026	4.54	1E-03	−43	−82	−2
Left VOTC[a]	25	0.043	−32.8	−44.8	−12.6	0.019	5.09	5E-04	−28	−40	−11
Right VOTC[a]	17	0.053	44.4	−52.0	−12.9	0.029	4.88	9E-04	44	−52	−11
Left mFG[a]	11	0.033	−45.5	20.8	30.2	0.036	3.85	2E-04	−46	23	28

Indented are MNI locations and statistics for peaks of additional local maxima within these clusters that were separated by at least 15 mm in the volume. The ROIs used in subsequent analyses were composed of ~100 voxels centered on the cross-format cluster peaks (*a*), with the addition of the local maximum for left premotor (ventral only; *b*). FWE, Familywise error rate; cor, corrected; unc, uncorrected.

Cross-format action category decoding across the brain

Our primary goal was to identify representations of action categories that were invariant to incidental visual elements, such as actors, objects, scene context, and the presence or absence of dynamic motion information. To this end, we scanned participants while they viewed videos and still images of eight categories of interactions. We then used a searchlight analysis to identify brain regions where action category representations could be decoded across the video and image formats. This analysis revealed seven contiguous clusters in the cross-format searchlight volume, which were located in left and right IPL, left and right lateral OTC (LOTC), left and right ventral OTC (VOTC), and left mFG (for a list of these clusters, see Fig. 2*A* and Table 1). These regions largely overlap with the previously identified AON (Caspers et al., 2010; Rizzolatti and Sinigaglia, 2010; Kilner, 2011; Urgesi et al., 2014). These results suggest that AON regions encode categories of actions in a consistent way across highly varied perceptual input. For subsequent analyses, ROIs corresponding to these clusters (~100 voxels each) were defined individually in each subject. (For discussion of the relationship of the cross-format OTC regions to functionally defined OTC regions based on previous literature, see below, "Cross-format decoding in functionally selective regions.")

The largest cluster, left IPL, had several local maxima (Table 1). The cluster peak was in left ventral IPL in the supramarginal gyrus ($x, y, z_{mni} = -58, -37, 28$) and this was used as the left IPL ROI for further analyses. An additional local maximum was located in left premotor cortex ($x, y, z_{mni} = -55, -4, 40$; Fig. 2*A*). Though this area was contiguous with the left IPL cluster in the volume, it is anatomically separated by several sulci and gyri from the other local maxima, and prior literature suggests a possible functionally distinct role for left premotor cortex in recognition of actions (Rizzolatti and Sinigaglia, 2010; Kilner, 2011; Caramazza et al., 2014; Wurm and Lingnau, 2015). Therefore, we defined an additional ROI around this local maximum for further interrogation. With this additional ROI, we had eight ROIs for subsequent analyses: left and right IPL, left and right LOTC, left and right VOTC, left premotor, and left mFG.

Prior work has shown coding of specific limb and effector information in some of the regions reported here (IPL and LOTC; Mahon et al., 2007; Peeters et al., 2009, 2013, Orlov et al., 2010, 2014, Gallivan et al., 2011, 2013a; Bracci and Peelen, 2013). To ensure that the present cross-decoding results were not driven solely by an effector-based distinction between action categories, we examined cross-format decoding separately for two sets of our action categories: those that involved hand/arm effectors (*massaging*, *pulling*, *shoving*, *slapping*, and *tapping*) versus those that involved other, more varied, effectors (*biting*, *brushing*, and *kicking*). If an effector-based distinction between hand/arm and non-hand/arm actions were driving our results, then we should observe cross-format decoding only within the varied effector set and not the hand/arm effector set. However, despite the reduced data available in each subset, we still observed cross-decoding in half or more of the cross-format ROIs in both subsets: Six of eight ROIs showed significant decoding within the varied effector set (left mFG, left VOTC, left and right LOTC, left and right IPL; *t* values >2.75, $p_{corrected}$ values <0.046) and 4 of 8 showed significant decoding within the hand/arm effector set (right VOTC, right LOTC, left and right IPL; *t* values >3.02, $p_{corrected}$ values <0.046). This suggests that, in these regions at least, cross-format decoding is unlikely to be driven solely by a coarse distinction between actions performed with the hand/arm versus other effectors.

Within-format action category decoding across the brain

To determine whether action category information tied to the particular visual format was present in other brain regions, we conducted whole-brain searchlights for action category decoding separately for the video format and image format. Within the video format, we found widespread action category decoding across the brain in both hemispheres (Fig. 2*B*). These results are not surprising given the consistency in visual motion energy across the video clips within action category (see above, Materials and Methods: "Video stimuli" section). In contrast, action category decoding within the image format was restricted largely to the regions identified in cross-format decoding (cf. Fig. 2*A*,*C*) with an additional left orbitofrontal cluster. This was confirmed in a conjunction overlap map of the three searchlight maps (Fig. 2*D*; Nichols et al., 2005): The within-format searchlights overlapped one another in or adjacent to areas observed in the cross-format searchlight. Interestingly, the degree of overlap of the maps in the different regions suggests a possible difference in the degree of format dependence of action coding between left IPL and the other regions. In the former, there is a large area of cross-decoding and the within-format territory (both image and video) overlaps with this. In the other regions, particularly the left LOTC, there is only a small area of cross-decoding, but large (and overlapping) areas of within-format decoding. This might suggest that action representations are less format-dependent in the left IPL than in other regions.

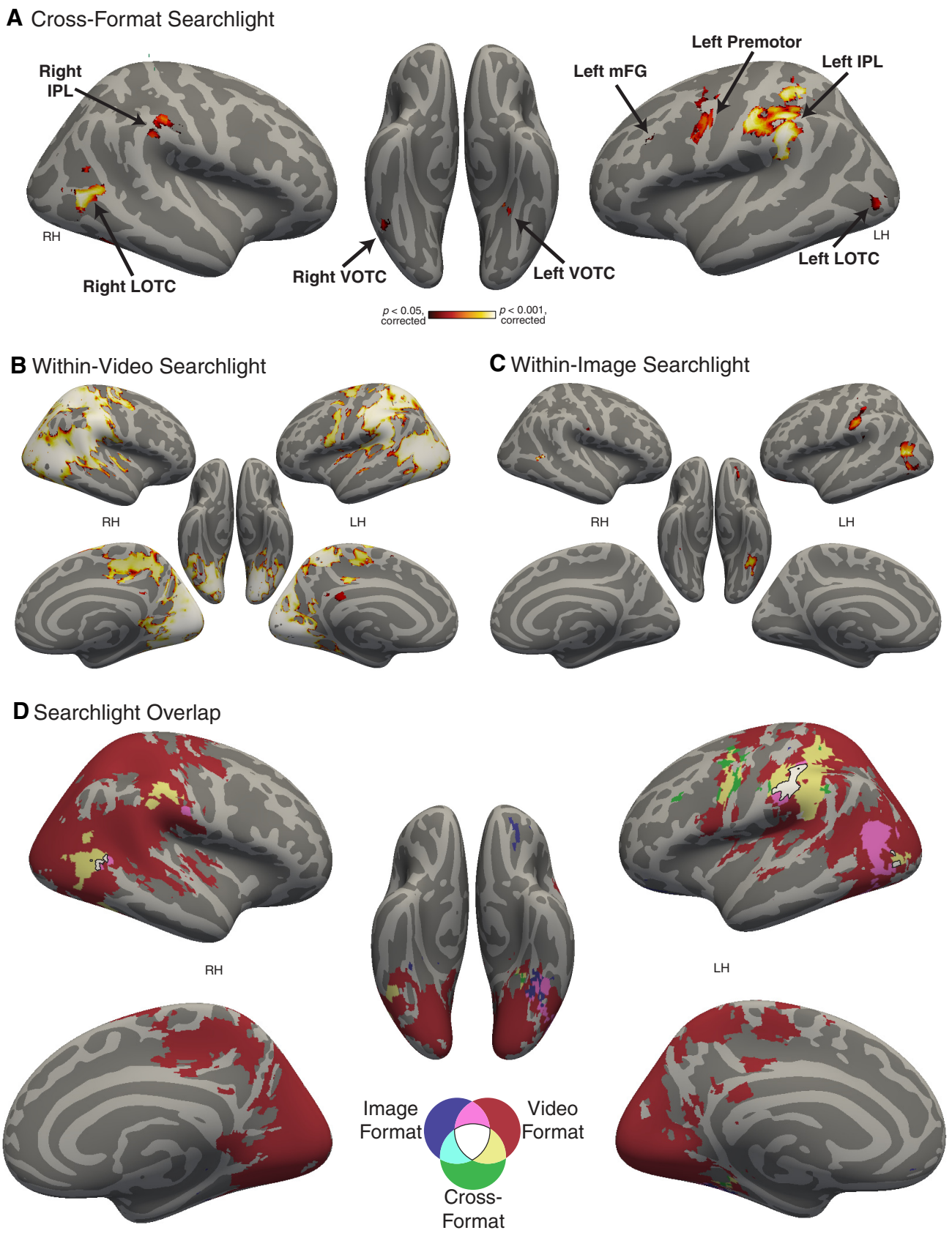

Figure 2. ***A***, Whole-brain searchlight for cross-format action category decoding. Black arrows and text indicate the anatomical locations of the cross-format clusters identified in this analysis, as well as the location of the ROI for left premotor cortex. Data were corrected for multiple comparisons at $p < 0.05$ using TFCE and permutation testing. For subsequent analyses, ROIs corresponding to these clusters were defined individually in each subject. ***B***, Whole-brain searchlight for Image Format action category decoding (corrected as in ***A***). ***C***, Whole-brain searchlight for Video Format action category decoding (corrected as in ***A***). ***D***, Conjunction overlap map of the searchlight analyses, with colors indicating which of the three searchlight maps overlap in each area (black outline indicates overlap of all three). Video Format decoding was widespread across the brain, whereas Image Format decoding was mostly restricted to similar regions as were found in the Cross-Format searchlight.

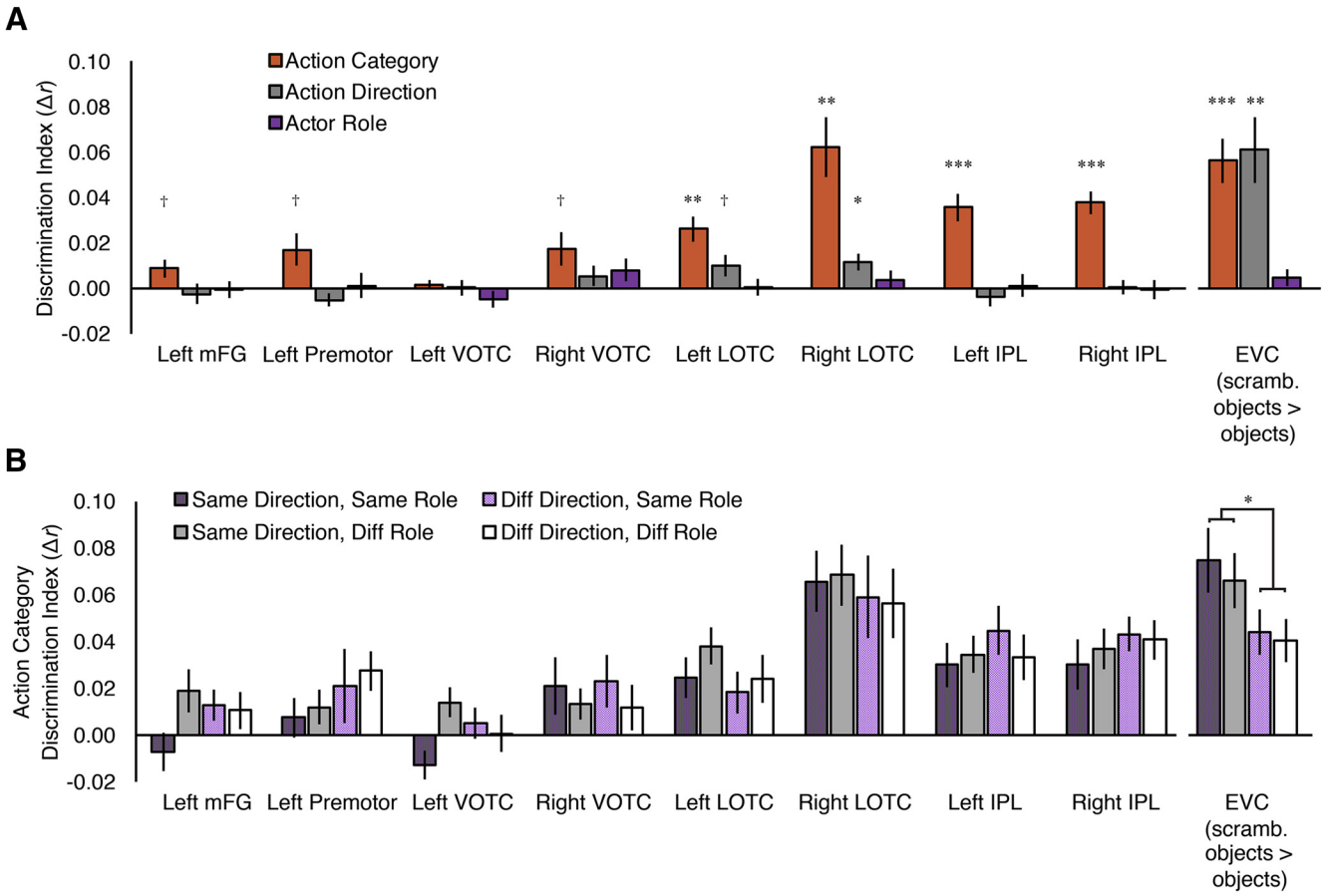

Figure 3. Analyses for action category specificity and generalization for the Video Format stimuli in cross-format ROIs. EVC, defined by a functional localizer as scrambled objects > intact objects, was also included for comparison with cross-format ROIs. ***A***, Decoding for action category, action direction, and actor roles. Discrimination index values shown here are average same minus average different correlation values for action category, action direction, and actor roles, respectively, collapsed over the other factors. Action direction could be decoded in left and right LOTC and EVC, whereas the actor role code could not be decoded in any regions. Action category could be decoded in most regions, though we note that this is necessitated by our ROI selection procedure, which was based on cross-format action category decoding using the same data. ***B***, Action category discrimination indices for the cross-format ROIs for each combination of action direction (same or different) and actor roles (same or different); that is, the orange Action Category bars in ***A*** split by the other factors. Only significant differences between action category decoding are indicated. Action category representations were largely invariant to the systematically manipulated properties of the video stimuli in cross-format ROIs, whereas in EVC, action category decoding was significantly better when action direction was the same versus different. †$p < 0.055$, uncorrected; *$p < 0.05$; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, corrected for multiple comparisons across the nine ROIs. Error bars indicate SEM.

Can the cross-format results be driven by similarities in spatial location of attention?

The cross-format results might have been trivially obtained if participants attended to similar spatial locations for each action category, even across the two visual formats (image and video). For example, it is reasonable to hypothesize that, for *kicking*, participants might have attended to the lower portion of the visual display, whereas for *slapping*, they attended to the higher portion. Such consistency in location of spatial attention has been shown to drive multivoxel responses in visual regions, including hMT+ (O'Connor et al., 2002; Bressler and Silver, 2010). To rule out this possibility, we conducted a control study in which a separate group of 16 participants performed the same task as the fMRI participants while their gaze was tracked with a remote eye tracker. These gaze data were analyzed similarly to how the fMRI data were analyzed; that is, multivariate patterns (here, 2D maps of gaze location) were constructed for each subject, format, and run, and discrimination indices were calculated from the correlation of the 2D gaze maps across action categories for each subject and format. If participants looked to consistent spatial locations for each action category, then these gaze map discrimination indices would be reliably above zero.

Action category could indeed be decoded based on gaze location for both the Image Format ($t_{(15)} = 2.60$, $p = 0.02$) and the Video format ($t_{(15)} = 7.91$, $p < 0.001$). However, across the visual formats, discrimination indices based on gaze locations were reliably below zero ($t_{(15)} = -4.26$, $p < 0.001$). These results indicate that gaze locations for action categories were consistent within format, but were systematically different across formats (e.g., looking at the top half of the screen for *kicking* in the Image Format, but the lower half for *kicking* in the Video Format). Therefore, absolute location of spatial attention is unlikely to explain the cross-format decoding results from fMRI data in the main experiment.

Invariance to systematically manipulated properties of the video stimuli

Abstract action category representations should show generalization, not only across formats, but also across variations in incidental properties within format such as actors or viewpoint/action direction. Some evidence that this may be the case comes from the fact that we were able to decode action category using only patterns elicited by the images even though the image stimuli were chosen to maximize within-category visual dissimilarity.

However, to formally test for generalization across incidental properties, we leveraged the fact that actor roles and action direction were systematically manipulated in the video clips. We extracted activity patterns within each ROI for each specific condition (i.e., each action category × action direction × actor role combination, 32 patterns in total). The correlation values between these conditions were then calculated and entered into repeated-measures ANOVAs (one for each ROI), with action category (same vs different), action direction (same vs different), and actor roles (same vs different) as factors. We also included EVC in the analysis (defined in a separate functional localizer as responses to scrambled objects > intact objects) as an indicator of whether it was possible to detect differences in action decoding across incidental low-level visual properties in our data.

Finding action category decoding was expected in this analysis because the ROIs were selected based on the presence of consistent action category patterns across format, which entails that the patterns within the Video Format should also be consistent. Somewhat surprisingly, action category decoding was robust in some, but not all, regions (Fig. 3*A*). This might be attributable to more variability in the estimates of activity patterns (there were two trials per β estimate in the GLM used here, as opposed to eight trials per β estimate in the previous analysis). Nevertheless, the estimates were consistent enough that seven of the eight cross-format ROIs, plus EVC, showed either a main effect of action category or a trend in this direction. These effects were marginal in left mFG, left premotor, and right VOTC ($F_{(1,14)}$ values of 4.66, 5.91, and 5.63, $p_{corrected}$ values = 0.12, $p_{uncorrected}$ values <0.05), and significant in all other regions ($F_{(1,14)}$ values >21.9, $p_{corrected}$ values <0.002, $p_{uncorrected}$ values <0.001) except left VOTC ($F_{(1,14)}$ = 0.75, $p_{corrected}$ = 0.40, $p_{uncorrected}$ = 0.40). For action direction, a subset of regions showed main effects (EVC and left LOTC significant, right LOTC marginal), with greater pattern similarity for the same action direction than different action direction (EVC: $F_{(1,14)}$ = 17.3, $p_{corrected}$ = 0.009, $p_{uncorrected}$ = 0.001; left LOTC: $F_{(1,14)}$ = 10.3, $p_{corrected}$ = 0.05, $p_{uncorrected}$ = 0.006; right LOTC: $F_{(1,14)}$ = 4.40, $p_{corrected}$ = 0.38, $p_{uncorrected}$ = 0.055; all other $F_{(1,14)}$ values <3.03, $p_{corrected}$ values >0.62, $p_{uncorrected}$ values >0.10; Fig. 3*A*). This suggests that these regions are sensitive to the direction of motion in the videos, which is not surprising given the presence of motion-selective regions (hMT+) in LOTC and the EVC's role in coding low-level visual features. No ROI showed a main effect of actor roles (all $F_{(1,14)}$ values <2.53, $p_{corrected}$ values >0.99, $p_{uncorrected}$ values >0.13; Fig. 3*A*), indicating that no region distinguished videos with Actor A as the agent from videos with Actor B as the agent.

Crucially, in terms of action category invariance, no cross-format ROI showed an interaction of action category with actor role and/or action direction; if anything, in left premotor cortex, action decoding was marginally better for different action direction versus same ($F_{(1,14)}$ = 4.07, $p_{corrected}$ = 0.51, $p_{uncorrected}$ = 0.06; all other cross-format ROI $F_{(1,14)}$ values <2.91, $p_{corrected}$ values >0.99, $p_{uncorrected}$ values >0.11; Fig. 3*B*). Although the lack of significant interactions is a null result and should be interpreted with caution, it is worthwhile to note that this modulation was detectable in our data: in EVC, action categories could be better decoded when action directions were the same than when they were different (action category × action direction interaction: $F_{(1,14)}$ = 12.4, $p_{corrected}$ = 0.03, $p_{uncorrected}$ = 0.003; Fig. 3*B*). Therefore, in regions showing cross-decoding of action category across videos and images, the ability to distinguish action categories was no greater when comparing across patterns elicited by videos in which actor role or action direction were the same than when comparing across videos in which actor role or action direction were different. Although we cannot rule out definitively the possibility that action representations in these cross-format regions are modulated by visual properties of the video stimuli, this finding is at least consistent with abstract action category codes.

Representational similarity analysis

Although recognizing actions as distinct from one another is a crucial first step toward action understanding, reasoning and communicating about actions requires more graded appreciation of similarities and differences between action categories. For example, two people may readily distinguish *slapping* from *shoving*, corresponding to successful action recognition for each individual. But if two people's representational spaces further indicate that *slapping* is very similar to *shoving*, mutual understanding and communication about these actions will be facilitated. To determine the extent to which representational spaces for action categories were consistent across individuals, we calculated the Spearman correlation between off-diagonal values of RDMs for each subject to every other subject separately for each cross-format ROI and comparison type (image format, video format, and cross-format). The mean intersubject correlation is the average consistency in representational spaces across individuals, where chance is zero.

For the image format comparisons, no ROI showed significant consistency in representational space across subjects ($p_{corrected}$ range 0.07–0.49), although four of the eight showed consistency uncorrected for multiple comparisons (left and right VOTC, right LOTC, and right IPL, $p_{uncorrected}$ < 0.05; other $p_{uncorrected}$ values >0.10). In contrast, for the video format, six of eight ROIs showed consistency across subjects (left premotor, right VOTC, left and right LOTC, and left and right IPL, $p_{corrected}$ values <0.05, $p_{uncorrected}$ values <0.005; other $p_{corrected}$ values >0.18, $p_{uncorrected}$ values >0.09). Similar findings to the video format were obtained for cross-format consistency: the same 6 ROIs showed consistency across subjects ($p_{corrected}$ values <0.03, $p_{uncorrected}$ values <0.009; other $p_{corrected}$ values >0.82, $p_{uncorrected}$ values >0.41). (Intersubject correlation values are depicted in Fig. 4*A*; see Fig. 4*B* for a visualization of the clustering of action categories across regions.)

It is at first glance puzzling that the cross-format consistency was reliable in most regions despite the lower image format consistency. One account of these contrasting results appeals to the difference in reliability of the "action category signal" between the image and video formats, which should be greater for the video format (as indicated by the higher norming label agreement in this format). For cross-format consistency, the robust video format action category signal may "pull out" the weaker image format signal even when the comparison is made across subjects. The plausibility of this account was confirmed by simulations. We generated sets of action category signals with different levels of signal-to-noise (SNR) and compared the resulting intersubject consistencies. Specifically, we generated one "true" activity pattern for each of eight action categories made up of 100 voxel responses randomly drawn from a Gaussian distribution $N(0,1)$. Varying degrees of noise were added to these "true" underlying action category patterns, separately for eight runs (four for each visual format) for each of 15 simulated subjects. This noise was varied systematically for each format by choosing SDs from the set (0.01, 0.10, 0.50, 1, 3) separately for the video and image formats (i.e., the video format SD might be 0.10, whereas the

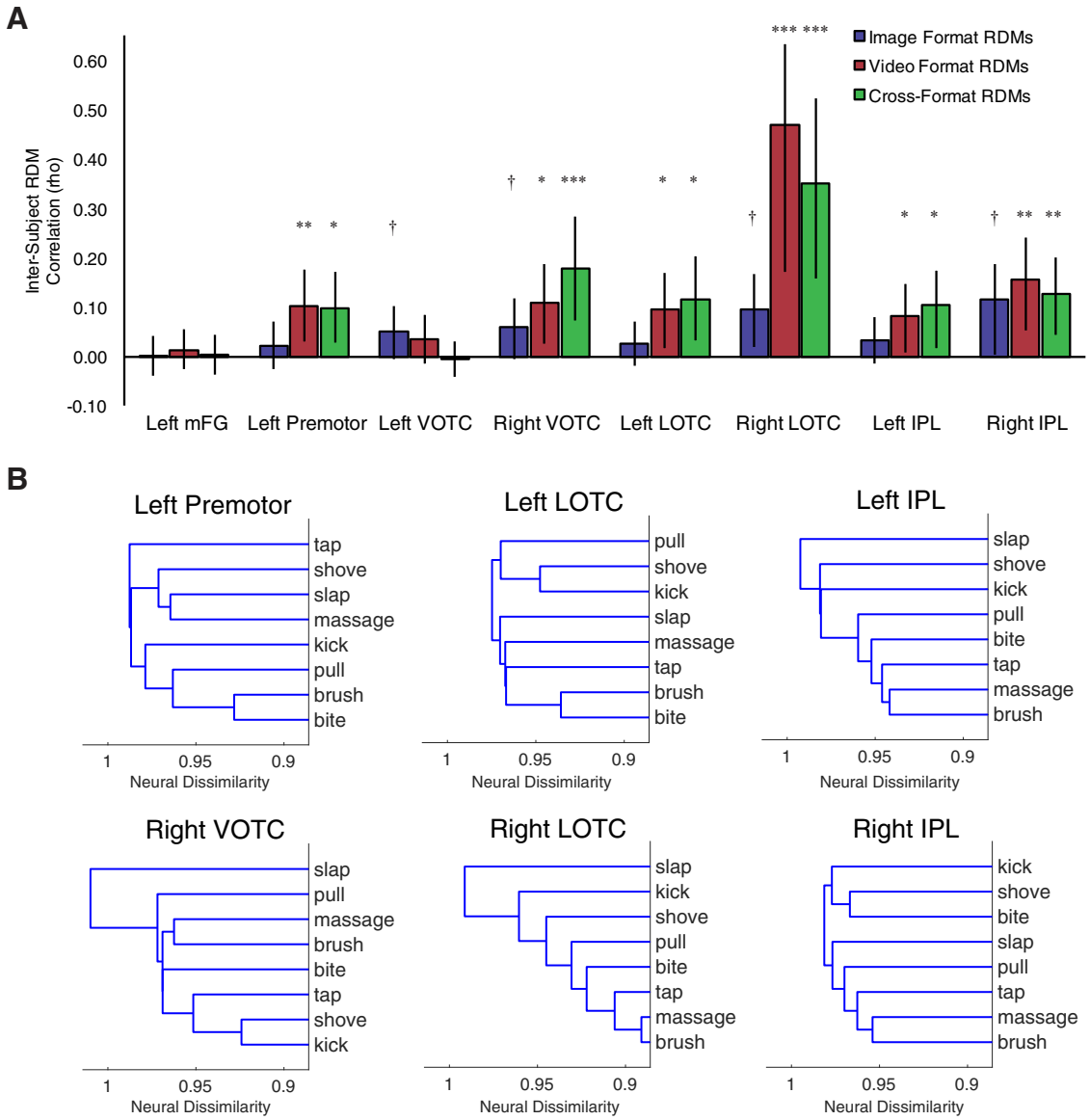

Figure 4. Cross-subject representational similarity analysis. RDMs for each subject were constructed from the multivoxel patterns for each action category and compared across subjects. ***A***, Mean intersubject RDM correlation across all pairwise comparisons of the 15 subjects separately for the image RDMs, video RDMs, and cross-format RDMs. Representational spaces for action categories were consistent across both subjects and formats in bilateral LOTC, bilateral IPL, right VOTC, and left premotor. †$p < 0.05$, uncorrected; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, corrected for multiple comparisons across the eight ROIs (separately for each comparison type). Permutation tests were used to determine significance based on a null distribution of the correlation statistic generated from 10,000 random permutations in which action category labels were shuffled before calculation of the RDM correlations. *P*-values are the proportion of permutation correlation statistics that were greater than the true statistic. Error bars here indicate the spread of the null distribution (95% of the null distribution width) centered at the mean intersubject RDM correlation value. ***B***, Dendrograms depicting the hierarchical clustering of action categories in each cross-format ROI that showed significant cross-format and cross-subject consistency based on mean cross-format RDMs across subjects. Neural dissimilarity is displayed in Pearson correlation distance ($1 - r$). Distances between clusters were computed using the MATLAB *linkage* function with the average distance algorithm.

image format SD might be 3). The same RSA as described above was then conducted using these simulated activity patterns. These simulations revealed that comparisons of RDMs built from two sets of low-SNR action category patterns (equivalent to the image-format comparison in this account) show a much less consistent relationship than comparisons of RDMs built from one high-SNR and one low-SNR set of action category patterns (the cross-format comparison in this account). Together, these analyses suggest that most of the regions that we have identified contain a representational space that generalizes from person to person, even when this space is built from two different visual formats.

Action category decoding in functionally selective regions

Although we focus above on regions identified in a hypothesis-free searchlight analysis, there are several well studied functional regions (fROIs) in or near to OTC that one might postulate a priori should have a role in action perception. These include motion-selective hMT+, body-selective EBA and FBA, object-selective LO and pFs, and pSTS-bio (Tootell et al., 1995; Kourtzi and Kanwisher, 2001; Grossman and Blake, 2002; Peelen et al., 2006; Kanwisher, 2010; Grill-Spector and Weiner, 2014). In addition, a region in LOTC just anterior to hMT+, the left pMTG, has been found to respond to linguistic descriptions of actions (Bedny et al., 2008, 2012, 2014; Peelen et al., 2012) and to respond

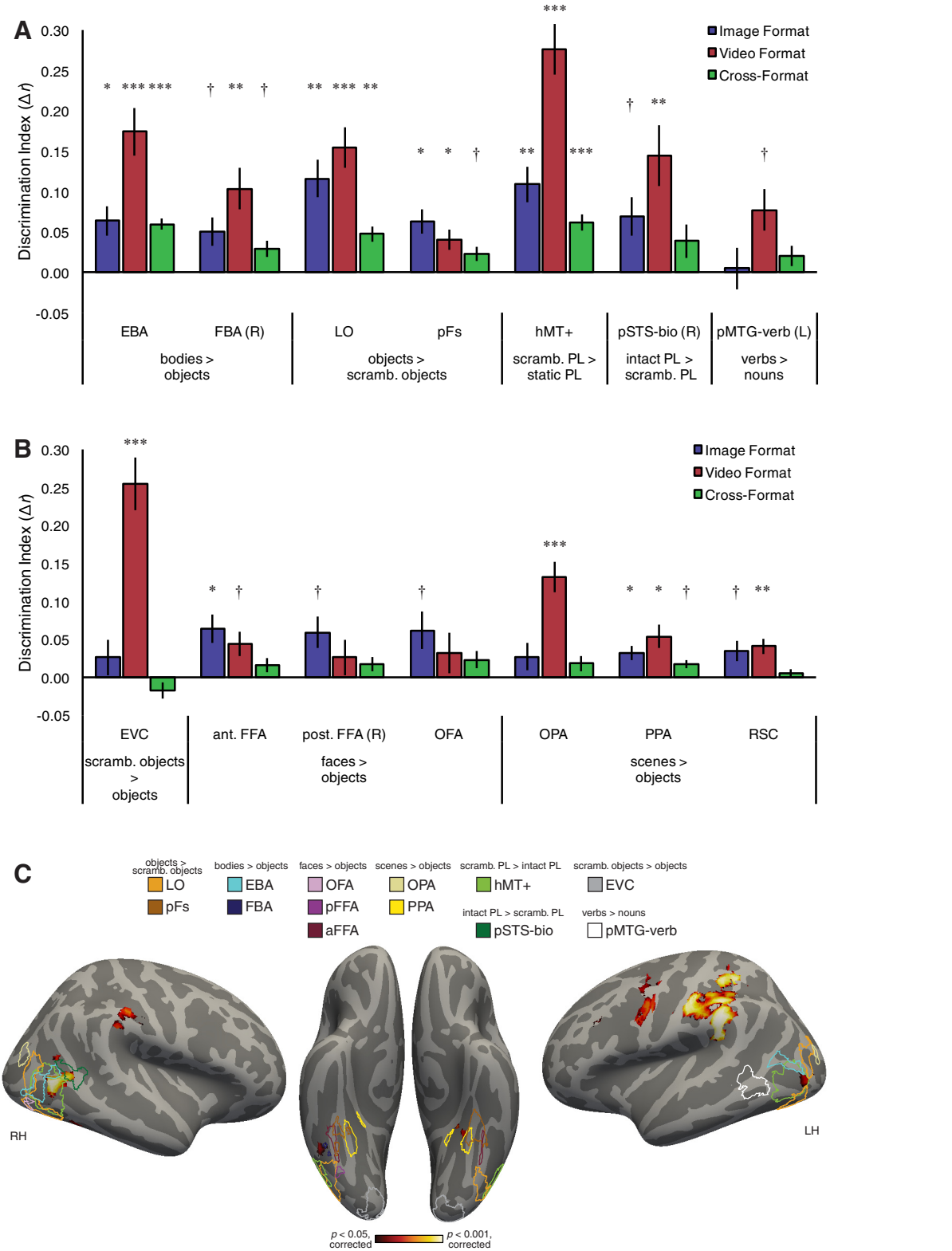

Figure 5. Action category discrimination indices for fROIs for each comparison type (within-image format, within-video format, and cross-format). The only fROIs tested in which significant cross-format decoding was found were EBA, LO, and hMT+. ***A***, Functional regions predicted to be sensitive to action category across format. ***B***, Functional regions predicted to show minimal sensitivity to action category across format. Listed below each fROI is the localizer contrast used to define the region (e.g., bodies > objects). †$p < 0.05$, uncorrected; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, corrected for multiple comparisons across the 14 fROIs (separately for each comparison type). Error bars indicate SEM. ***C***, Visualization of the locations of fROIs relative to brain regions in which cross-format action category decoding was found. Cross-decoding searchlight map is identical to that in Figure 2*A* (i.e., corrected for multiple comparisons at $p < 0.05$). Outlines of fROIs were created by transforming individual subjects' fROIs to standard space and computing a group *t* statistic. Group fROIs were thresholded at $p < 0.001$ (uncorrected) except for the following fROIs, for which lower thresholds were needed for visualization: Left OPA, right OFA, and pMTG-verb ($p < 0.01$); left EBA ($p < 0.05$); and left anterior FFA, right FBA, and right posterior FFA ($p < 0.33$). RSC is not shown because no significant cross-decoding appeared on medial surfaces. ant., Anterior; aFFA and pFFA, anterior and posterior FFA, respectively; L, left; PL, point-light; post., posterior; R, right; scramb., scrambled.

in action tasks involving both words and static images (Watson et al., 2013). To test the possibility that some of these regions might support abstract action representations, we performed the cross-format decoding analysis in these fROIs (see also Gallivan et al., 2013b; Gallivan and Culham, 2015). As a control, we also examined other fROIs that we did not expect to be involved in abstract action category representations (face- and scene-selective regions and early visual cortex; Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Kanwisher, 2010).

The only fROIs tested in which significant cross-format decoding was found were EBA, LO, and hMT+ ($t_{(14)}$ values >5.20, $p_{corrected}$ values <0.002, $p_{uncorrected}$ values <0.001; Fig. 5*A*,*B*). We did not find evidence for reliable cross-format decoding in other regions, although FBA, pFs, and PPA showed cross-format decoding at an uncorrected level (t values of 2.98, 2.78, and 3.02, respectively, $p_{corrected}$ values <0.13, $p_{uncorrected}$ values <0.02; all other t values <2.13, $p_{corrected}$ values >0.41, $p_{uncorrected}$ values >0.052). Notably, we did not find clear evidence for cross-format action category decoding in two regions known to code for action-relevant stimuli: right pSTS-bio ($t_{(10)} = 1.92$, $p_{corrected} = 0.59$, $p_{uncorrected} = 0.08$) and left pMTG-verb ($t_{(10)} = 1.66$, $p_{corrected} = 0.59$, $p_{uncorrected} = 0.13$). Together, these results suggest that the EBA, LO, and hMT+ are not only involved in representing bodies, objects, and motion, but also contribute to analysis of visual action scenes at an abstract level. (For a qualitative sense of the spatial relationship of fROIs and cross-format searchlight decoding, see Fig. 5*C*.)

In addition to the cross-format results in EBA, hMT+, and LO, several fROIs were sensitive to the action category information depicted within only the video format or image format (Fig. 5*A*,*B*). However, the fact that these fROIs did not demonstrate cross-format decoding suggests that their role in representing actions at an abstract level is limited. Furthermore, the absence of within-image format decoding in early visual cortex suggests that we adequately varied low-level image properties within action category.

Discussion

The goal of this study was to identify brain regions that mediate visual recognition of actions. We posited that these regions should display three key properties. First, they should support representations that discriminate between action categories, but are at least partially invariant to incidental features such as actor role, scene background, or viewpoint. Second, these action representations should be elicitable by both dynamic and static perceptual input. Third, these regions should not only discriminate hand-object interactions, but also whole-body interactions with different effectors. By using cross-format decoding methods, we identified several regions with these properties: bilateral OTC (lateral and ventral), bilateral IPL, left premotor cortex, and left mFG. The subset of these regions previously identified as the AON (LOTC, IPL, and left premotor; Caspers et al., 2010; Rizzolatti and Sinigaglia, 2010; Kilner, 2011; Urgesi et al., 2014) also exhibited consistency in representational space across subjects, a property that can facilitate a common understanding of actions among individuals.

Our findings add to the growing evidence that LOTC is involved in the coding of action categories (Oosterhof et al., 2010, 2012a, 2012b; Gallivan et al., 2013b; Watson et al., 2013; Gallivan and Culham, 2015; Tarhan et al., 2015; Tucciarelli et al., 2015; Wurm and Lingnau, 2015; Wurm et al., 2016; for review, see Lingnau and Downing, 2015). In particular, our analyses of functional ROIs indicated that areas in LOTC selective for bodies, objects, and motion are also involved in visual action recognition from varied perceptual input: cross-format action category decoding was observed in EBA, LO, and hMT+ (see above, "Action category decoding in functionally selective regions" section; Downing et al., 2001; Kourtzi and Kanwisher, 2001; Peelen et al., 2006; Ferri et al., 2013; Weiner and Grill-Spector, 2013; for review, see Lingnau and Downing, 2015). In contrast, we failed to observe cross-format decoding in several functionally defined regions known to be responsive to action-relevant stimuli: left pMTG-verb (Bedny et al., 2008, 2012, 2014; Peelen et al., 2012; Watson et al., 2013) and pSTS-bio (Vaina et al., 2001; Grossman and Blake, 2002; Peuskens et al., 2005; Gao et al., 2012; Deen et al., 2015). Although this latter set of null results should be interpreted with caution, it suggests that these regions might be involved in processing the lexical semantics of actions (pMTG-verb) or the motion of animate entities (pSTS-bio) rather than being involved in recognition of visual action categories per se.

Our results also agree with work suggesting that IPL is involved in abstract coding of actions. IPL has been implicated in the representation of dynamic upper-limb actions (Cattaneo et al., 2010; Bach et al., 2010; Abdollahi et al., 2013; Ferri et al., 2015) and tool-related actions (Mahon et al., 2007; Peeters et al., 2009, 2013; Gallivan et al., 2011; Tarhan et al., 2015; for review, see Orban and Caruana, 2014; Gallivan and Culham, 2015). Other work suggests that IPL, particularly in the left hemisphere, may represent the abstract causal outcomes or relationships between entities. For example, Oosterhof et al. (2012b) found crossmodal action-specific codes across execution and mental imagery in left IPL, but not in premotor cortex or LOTC. Left IPL exhibits adaptation when viewing reaching actions toward the same goal object, even when the hand follows a very different spatial trajectory (Hamilton and Grafton, 2006). Moreover, activation patterns in left IPL have been found to distinguish between motor acts (e.g., *pushing*, *grasping*), but generalize across acts performed with different body parts (Jastorff et al., 2010). Recent work from Leshinskaya and Caramazza (2015) suggests that a dorsal portion of left IPL represents common outcomes associated with different objects even when those outcomes are defined at a highly abstract level (e.g., a wind chime for decorating a house and perfume for decorating oneself). In our study, the spatial extent of cross-decoding was greater in the left hemisphere than the right (Fig. 2, Table 1). Together, previous work and the current study suggest a role for IPL (particularly on the left) in representation of actions at an abstract level.

We also observed action decoding in premotor cortex. Like LOTC and IPL, premotor cortex has been consistently implicated in action observation (Buccino et al., 2004; Gazzola et al., 2007; Saygin, 2007; Etzel et al., 2008; Majdandzic et al., 2009; Ogawa and Inui, 2011; for a meta-analysis, see Caspers et al., 2010), a finding that has been taken to support motor theories of action understanding (e.g., Rizzolatti and Craighero, 2004; Rizzolatti and Sinigaglia, 2010). In contrast, cognitive theories maintain that action understanding is achieved via higher-level, amodal representations (Hickok, 2009; Caramazza et al., 2014). Because our study only examines action observation, not action execution, we cannot address the crossmodal (observe/execute) aspects of this debate (Chong et al., 2008; Dinstein et al., 2008; Kilner et al., 2009; Caspers et al., 2010; Oosterhof et al., 2010, 2012a; Tarhan et al., 2015). Nevertheless, we did find that, along with LOTC and IPL, representations of observed actions in premotor cortex were invariant to incidental perceptual features and the dynamicity of visual input. Although this result might seem superficially at odds with Wurm and Lingnau's (2015) finding that represen-

tations of *open* and *close* generalized across the acted-upon object and the associated action kinematics in IPL and LOTC but not in premotor cortex, we believe that our result is not necessarily inconsistent. Whereas Wurm and Lingnau (2015) defined their actions by object state changes (*open* vs *close*), we defined our actions by the physical manner of interaction (e.g., *kick* vs *massage*). These components are logically dissociable (e.g., one can kick a door open or closed). Therefore, AON regions may differ in which components of actions they represent, with premotor coding for the physical manner of action but not state change and LOTC and IPL coding for both. In any case, our results support the idea that there is abstraction across some features of perceptual input in all AON regions, including premotor cortex.

An open question is how the AON can extract common action codes from both static and dynamic displays. Given that in naturalistic action observation, all body parts of actors are generally visible, simple presence/absence of specific effectors in the visual field cannot be sufficient for recognition. Instead, we hypothesize that the spatial configuration of entities (actor/effector and acted-upon entity) is crucial for determining the action category and that parts of the AON process this configural information. Such information would be observable in both images and videos. Supporting this view, there is behavioral and neuroimaging evidence that the visual system codes the elements of actions as a perceptual unit, possibly including information about their spatial configuration, rather than simply coding them as separate, distinct items. First, briefly observed snapshots of actions are sufficient for recognition, but only when the configuration of scene entities is consistent with the given action (Dobel et al., 2007; Hafri et al., 2013). Second, multivoxel patterns in LOTC elicited by images of interacting humans and objects are not linearly decodable from the patterns elicited by the same actors and objects shown in isolation, yet such linear decoding is successful if the actor and objects are superimposed in a noninteracting manner (Baldassano et al., 2016). This suggests that neural representations of human–object interactions (at least in LOTC) may incorporate configuration information that makes them more than the sum of their visual parts.

Another possible explanation for common static/dynamic action codes, not mutually exclusive to the above, is that, through experience, static snapshots of actions become associated with full action sequences and thus elicit those sequences (Giese and Poggio, 2003; Jastorff et al., 2009; Singer and Sheinberg, 2010; Vangeneugden et al., 2011). This association may account for the implicit/implied motion effects observed in both behavioral and neuroimaging studies (Freyd, 1983; Shiffrar and Freyd, 1993; Kourtzi and Kanwisher, 2000; Senior et al., 2000; Winawer et al., 2008, 2010; Gervais et al., 2010) and may be what allows the action recognition system to be robust to missing or ambiguous perceptual input. Supporting this idea, behavioral work has shown that causal representations are engaged for both simple and naturalistic launching events despite temporary occlusion or absence of the causal moment from the stimulus display (Strickland and Keil, 2011; Bae and Flombaum, 2011).

To summarize, we uncovered abstract neural codes for action categories in bilateral OTC and IPL, left premotor cortex, and left mFG, including regions of LOTC that have been previously implicated in body, object, and motion processing. These codes were invariant to differences in actors, objects, scene context, or viewpoint and could be evoked by both dynamic and static stimuli. Moreover, most of these regions showed consistent representational spaces across subjects and formats, which is a feature of an action recognition system that can facilitate a common understanding of actions across individuals. Together, our findings suggest that these regions mediate abstract representations of actions that may provide a link between visual systems that support perceptual recognition of actions and conceptual systems that support flexible, complex thought about physically interacting entities.

References

Abdollahi RO, Jastorff J, Orban GA (2013) Common and segregated processing of observed actions in human SPL. Cereb Cortex 23:2734–2753. CrossRef Medline

Bach P, Peelen MV, Tipper SP (2010) On the role of object information in action observation: an fMRI study. Cereb Cortex 20:2798–2809. CrossRef Medline

Bae GY, Flombaum JI (2011) Amodal causal capture in the tunnel effect. Perception 40:74–90. CrossRef Medline

Baldassano C, Beck DM, Fei-Fei L (2016) Human-object interactions are more than the sum of their parts. Cereb Cortex. Advance online publication. Retrieved February 23, 2017. doi:10.1093/cercor/bhw077.

Bedny M, Caramazza A, Grossman E, Pascual-Leone A, Saxe R (2008) Concepts are more than percepts: the case of action verbs. J Neurosci 28:11347–11353. CrossRef Medline

Bedny M, Caramazza A, Pascual-Leone A, Saxe R (2012) Typical neural representations of action verbs develop without vision. Cereb Cortex 22:286–293. CrossRef Medline

Bedny M, Dravida S, Saxe R (2014) Shindigs, brunches, and rodeos: the neural basis of event words. Cogn Affect Behav Neurosci 14:891–901. CrossRef Medline

Bracci S, Peelen MV (2013) Body and object effectors: the organization of object representations in high-level visual cortex reflects body-object interactions. J Neurosci 33:18247–18258. CrossRef Medline

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436. CrossRef Medline

Bressler DW, Silver MA (2010) Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. Neuroimage 53:526–533. CrossRef Medline

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

Caramazza A, Anzellotti S, Strnad L, Lingnau A (2014) Embodied cognition and mirror neurons: a critical assessment. Annu Rev Neurosci 37:1–15. CrossRef Medline

Caspers S, Zilles K, Laird AR, Eickhoff SB (2010) ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50:1148–1167. CrossRef Medline

Cattaneo L, Sandrini M, Schwarzbach J (2010) State-dependent TMS reveals a hierarchical representation of observed acts in the temporal, parietal, and premotor cortices. Cereb Cortex 20:2252–2258. CrossRef Medline

Chong TT, Cunnington R, Williams MA, Kanwisher N, Mattingley JB (2008) fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Curr Biol 18:1576–1580. CrossRef Medline

O'Connor DH, Fukui MM, Pinsk MA, Kastner S (2002) Attention modulates responses in the human lateral geniculate nucleus. Nat Neurosci 5:1203–1209. CrossRef Medline

Deen B, Koldewyn K, Kanwisher N, Saxe R (2015) Functional organization of social perception and cognition in the superior temporal sulcus. Cereb Cortex 25:4596–4609. CrossRef Medline

Dinstein I, Gardner JL, Jazayeri M, Heeger DJ (2008) Executed and observed movements have different distributed representations in human aIPS. J Neurosci 28:11231–11239. CrossRef Medline

Dobel C, Gumnior H, Bölte J, Zwitserlood P (2007) Describing scenes hardly seen. Acta Psychol (Amst) 125:129–143. CrossRef Medline

Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. Science 293:2470–2473. CrossRef Medline

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601. CrossRef Medline

Etzel JA, Gazzola V, Keysers C (2008) Testing simulation theory with cross-modal multivariate classification of fMRI data. PLoS One 3:e3690. CrossRef Medline

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558. CrossRef Medline
Ferri S, Kolster H, Jastorff J, Orban GA (2013) The overlap of the EBA and the MT/V5 cluster. Neuroimage 66:412–425. CrossRef Medline
Ferri S, Rizzolatti G, Orban GA (2015) The organization of the posterior parietal cortex devoted to upper limb actions: an fMRI study. Hum Brain Mapp 36:3845–3866. CrossRef Medline
Fischl B, Sereno MI, Tootell RB, Dale AM (1999) High-resolution intersubject averaging and a surface-based coordinate system. Hum Brain Mapp 8:272–284. CrossRef Medline
Freyd JJ (1983) The mental representation of movement when static stimuli are viewed. Percept Psychophys 33:575–581. CrossRef Medline
Gallivan JP, Culham JC (2015) Neural coding within human brain areas involved in actions. Curr Opin Neurobiol 33:141–149. CrossRef Medline
Gallivan JP, McLean DA, Smith FW, Culham JC (2011) Decoding effector-dependent and effector-independent movement intentions from human parieto-frontal brain activity. J Neurosci 31:17149–17168. CrossRef Medline
Gallivan JP, Adam McLean D, Valyear KF, Culham JC (2013a) Decoding the neural mechanisms of human tool use. Elife 2013:2:e00425. CrossRef Medline
Gallivan JP, Chapman CS, McLean DA, Flanagan JR, Culham JC (2013b) Activity patterns in the category-selective occipitotemporal cortex predict upcoming motor actions. Eur J Neurosci 38:2408–2424. CrossRef Medline
Gao T, Scholl BJ, McCarthy G (2012) Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. J Neurosci 32:14276–14280. CrossRef Medline
Gazzola V, van der Worp H, Mulder T, Wicker B, Rizzolatti G, Keysers C (2007) Aplasics born without hands mirror the goal of hand actions with their feet. Curr Biol 17:1235–1240. CrossRef Medline
Gervais WM, Reed CL, Beall PM, Roberts RJ Jr (2010) Implied body action directs spatial attention. Atten Percept Psychophys 72:1437–1443. CrossRef Medline
Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4:179–192. CrossRef Medline
Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci 15: 536–548. CrossRef Medline
Grossman ED, Blake R (2002) Brain areas active during visual perception of biological motion. Neuron 35:1167–1175. CrossRef Medline
Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R (2000) Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711–720. CrossRef Medline
Hafri A, Papafragou A, Trueswell JC (2013) Getting the gist of events: Recognition of two-participant actions from brief displays. J Exp Psychol Gen 142:880–905. CrossRef Medline
Hamilton AF, Grafton ST (2006) Goal representation in human anterior intraparietal sulcus. J Neurosci 26:1133–1137. CrossRef Medline
Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430. CrossRef Medline
Hickok G (2009) Eight problems for the mirror neuron theory of action understanding in monkeys and humans. J Cogn Neurosci 21:1229–1243. CrossRef Medline
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70.
Jastorff J, Kourtzi Z, Giese MA (2009) Visual learning shapes the processing of complex movement stimuli in the human brain. J Neurosci 29:14026–14038. CrossRef Medline
Jastorff J, Begliomini C, Fabbri-Destro M, Rizzolatti G, Orban GA (2010) Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. J Neurophysiol 104:128–140. CrossRef Medline
Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841. CrossRef Medline
Julian JB, Fedorenko E, Webster J, Kanwisher N (2012) An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60:2357–2364. CrossRef Medline
Kable JW, Chatterjee A (2006) Specificity of action representations in the lateral occipitotemporal cortex. J Cogn Neurosci 18:1498–1517. CrossRef Medline
Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. Proc Natl Acad Sci U S A 107:11163–11170. CrossRef Medline
Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311. Medline
Kilner JM (2011) More than one pathway to action understanding. Trends Cogn Sci 15:352–357. CrossRef Medline
Kilner JM, Neal A, Weiskopf N, Friston KJ, Frith CD (2009) Evidence of mirror neurons in human inferior frontal gyrus. J Neurosci 29:10153–10159. CrossRef Medline
Kourtzi Z, Kanwisher N (2000) Activation in human MT/MST by static images with implied motion. J Cogn Neurosci 12:48–55. CrossRef Medline
Kourtzi Z, Kanwisher N (2001) Representation of perceived object shape by the human lateral occipital complex. Science 293:1506–1509. CrossRef Medline
Kriegeskorte N, Kievit RA (2013) Representational geometry: Integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412. CrossRef Medline
Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline
Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 2:4. CrossRef Medline
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540. CrossRef Medline
Leshinskaya A, Caramazza A (2015) Abstract categories of functions in anterior parietal lobe. Neuropsychologia 76:27–40. CrossRef Medline
Lingnau A, Downing PE (2015) The lateral occipitotemporal cortex in action. Trends Cogn Sci 19:268–277. CrossRef Medline
Mahon BZ, Milleville SC, Negri GA, Rumiati RI, Caramazza A, Martin A (2007) Action-related properties shape object representations in the ventral stream. Neuron 55:507–520. CrossRef Medline
Majdandzic J, Bekkering H, van Schie HT, Toni I (2009) Movement-specific repetition suppression in ventral and dorsal premotor cortex during action observation. Cereb Cortex 19:2736–2745. CrossRef Medline
Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25. CrossRef Medline
Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660. CrossRef Medline
Ogawa K, Inui T (2011) Neural representation of observed actions in the parietal and premotor cortex. Neuroimage 56:728–735. CrossRef Medline
Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis 42:145–175. CrossRef
Oosterhof NN, Wiggett AJ, Diedrichsen J, Tipper SP, Downing PE (2010) Surface-based information mapping reveals crossmodal vision–action representations in human parietal and occipitotemporal cortex. J Neurophysiol 104:1077–1089. CrossRef Medline
Oosterhof NN, Tipper SP, Downing PE (2012a) Viewpoint (in)dependence of action representations: an MVPA study. J Cogn Neurosci 24:975–989. CrossRef Medline
Oosterhof NN, Tipper SP, Downing PE (2012b) Visuo-motor imagery of specific manual actions: a multi-variate pattern analysis fMRI study. Neuroimage 63:262–271. CrossRef Medline
Oosterhof NN, Tipper SP, Downing PE (2013) Crossmodal and action-specific: neuroimaging the human mirror neuron system. Trends Cogn Sci 17:311–318. CrossRef Medline
Orban GA, Caruana F (2014) The neural basis of human tool use. Front Psychol 5:310. CrossRef Medline
Orlov T, Makin TR, Zohary E (2010) Topographic representation of the human body in the occipitotemporal cortex. Neuron 68:586–600. CrossRef Medline
Orlov T, Porat Y, Makin TR, Zohary E (2014) Hands in motion: an upper-

limb-selective area in the occipitotemporal cortex shows sensitivity to viewed hand kinematics. J Neurosci 34:4882–4895. CrossRef Medline
Peelen MV, Wiggett AJ, Downing PE (2006) Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. Neuron 49:815–822. CrossRef Medline
Peelen MV, Romagno D, Caramazza A (2012) Independent representations of verbs and actions in left lateral temporal cortex. J Cogn Neurosci 24: 2096–2107. CrossRef Medline
Peeters RR, Rizzolatti G, Orban GA (2013) Functional properties of the left parietal tool use region. Neuroimage 78:83–93. CrossRef Medline
Peeters R, Simone L, Nelissen K, Fabbri-Destro M, Vanduffel W, Rizzolatti G, Orban GA (2009) The representation of tool use in humans and monkeys: Common and uniquely human features. J Neurosci 29:11523–11539. CrossRef Medline
Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442. CrossRef Medline
Peuskens H, Vanrie J, Verfaillie K, Orban GA (2005) Specificity of regions processing biological motion. Eur J Neurosci 21:2864–2875. CrossRef Medline
Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192. CrossRef Medline
Rizzolatti G, Sinigaglia C (2010) The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat Rev Neurosci 11:264–274. CrossRef Medline
Saygin AP (2007) Superior temporal and premotor brain areas necessary for biological motion perception. Brain 130:2452–2461. CrossRef Medline
Schwarzlose RF, Swisher JD, Dang S, Kanwisher N (2008) The distribution of category and location information across object-selective regions in human visual cortex. Proc Natl Acad Sci U S A 105:4447–4452. CrossRef Medline
Senior C, Barnes J, Giampietro V, Simmons A, Bullmore ET, Brammer M, David AS (2000) The functional neuroanatomy of implicit-motion perception or representational momentum. Curr Biol 10:16–22. Medline
Shiffrar M, Freyd J (1993) Timing and apparent motion path choice with human body photographs. Psychol Sci 4:379–384. CrossRef
Singer JM, Sheinberg DL (2010) Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J Neurosci 30:3133–3145. CrossRef Medline
Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83–98. CrossRef Medline
Strickland B, Keil F (2011) Event completion: event based inferences distort memory in a matter of seconds. Cognition 121:409–415. CrossRef Medline
Tarhan LY, Watson CE, Buxbaum LJ (2015) Shared and distinct neuroanatomic regions critical for tool-related action production and recognition: evidence from 131 left-hemisphere stroke patients. J Cogn Neurosci 27: 2491–2511. CrossRef Medline
Tootell RB, Reppas JB, Kwong KK, Malach R, Born RT, Brady TJ, Rosen BR, Belliveau JW (1995) Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. J Neurosci 15:3215–3230. Medline
Tucciarelli R, Turella L, Oosterhof NN, Weisz N, Lingnau A (2015) MEG multivariate analysis reveals early abstract action representations in the lateral occipitotemporal cortex. J Neurosci 35:16034–16045. CrossRef Medline
Urgesi C, Candidi M, Avenanti A (2014) Neuroanatomical substrates of action perception and understanding: an anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain injured patients. Front Hum Neurosci 8:344. CrossRef Medline
Vaina LM, Solomon J, Chowdhury S, Sinha P, Belliveau JW (2001) Functional neuroanatomy of biological motion perception in humans. Proc Natl Acad Sci U S A 98:11656–11661. CrossRef Medline
van Boxtel JJA, Lu H (2013) A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. J Vis 13: pii: 7. CrossRef Medline
Vangeneugden J, De Mazière PA, Van Hulle MM, Jaeggli T, Van Gool L, Vogels R (2011) Distinct mechanisms for coding of visual actions in macaque temporal cortex. J Neurosci 31:385–401. CrossRef Medline
Vangeneugden J, Peelen MV, Tadin D, Battelli L (2014) Distinct neural mechanisms for body form and body motion discriminations. J Neurosci 34:574–585. CrossRef Medline
Watson CE, Cardillo ER, Bromberger B, Chatterjee A (2014) The specificity of action knowledge in sensory and motor systems. Front Psychol 5:494. CrossRef Medline
Watson CE, Cardillo ER, Ianni GR, Chatterjee A (2013) Action concepts in the brain: an activation likelihood estimation meta-analysis. J Cogn Neurosci 25:1191–1205. CrossRef Medline
Weiner KS, Grill-Spector K (2013) Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. Psychol Res 77:74–97. CrossRef Medline
Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE toolbox. Behav Res Methods 42:671–684. CrossRef Medline
Winawer J, Huk AC, Boroditsky L (2008) A motion aftereffect from still motion photographs depicting motion. Psychol Sci 19:276–283. CrossRef Medline
Winawer J, Huk AC, Boroditsky L (2010) A motion aftereffect from visual imagery of motion. Cognition 114:276–284. CrossRef Medline
Wurm MF, Lingnau A (2015) Decoding actions at different levels of abstraction. J Neurosci 35:7727–7735. CrossRef Medline
Wurm MF, Ariani G, Greenlee MW, Lingnau A (2016) Decoding concrete and abstract action representations during explicit and implicit conceptual processing. Cereb Cortex 26:3390–3401. CrossRef Medline