Behavioral/Cognitive

Disentangling Object Category Representations Driven by Dynamic and Static Visual Input

Sophia Robert,[1,2] **Leslie G. Ungerleider,**[1] **and Maryam Vaziri-Pashkam**[1]
[1]Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, Maryland 20892, and [2]Department of Psychology and Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Humans can label and categorize objects in a visual scene with high accuracy and speed, a capacity well characterized with studies using static images. However, motion is another cue that could be used by the visual system to classify objects. To determine how motion-defined object category information is processed by the brain in the absence of luminance-defined form information, we created a novel stimulus set of "object kinematograms" to isolate motion-defined signals from other sources of visual information. Object kinematograms were generated by extracting motion information from videos of 6 object categories and applying the motion to limited-lifetime random dot patterns. Using functional magnetic resonance imaging (fMRI) ($n = 15$, 40% women), we investigated whether category information from the object kinematograms could be decoded within the occipitotemporal and parietal cortex and evaluated whether the information overlapped with category responses to static images from the original videos. We decoded object category for both stimulus formats in all higher-order regions of interest (ROIs). More posterior occipitotemporal and ventral regions showed higher accuracy in the static condition, while more anterior occipitotemporal and dorsal regions showed higher accuracy in the dynamic condition. Further, decoding across the two stimulus formats was possible in all regions. These results demonstrate that motion cues can elicit widespread and robust category responses on par with those elicited by static luminance cues, even in ventral regions of visual cortex that have traditionally been associated with primarily image-defined form processing.

Key words: **biological motion; dynamic; motion-defined shape; object category; object information; static**

Significance Statement

Much research on visual object recognition has focused on recognizing objects in static images. However, motion is a rich source of information that humans might also use to categorize objects. Here, we present the first study to compare neural representations of several animate and inanimate objects when category information is presented in two formats: static cues or isolated dynamic motion cues. Our study shows that, while higher-order brain regions differentially process object categories depending on format, they also contain robust, abstract category representations that generalize across format. These results expand our previous understanding of motion-derived animate and inanimate object category processing and provide useful tools for future research on object category processing driven by multiple sources of visual information.

Introduction

Humans can categorize objects with striking speed and accuracy. Previous research on the neural basis of visual object recognition has focused on the processing of static features from images along the ventral hierarchy of the primate brain (for review, see Peissig and Tarr, 2007). However, real-world scenes are not static. Indeed, motion cues can contain category-relevant information that humans use to make judgments about objects. Behavioral studies using point-light displays (PLDs) (Johansson, 1973, 1976) have established that, even with the impoverished motion information available in PLDs, humans can quickly perceive a moving person, identify the action being performed, and even determine the actor's age, gender, and affect (e.g., Cutting and Kozlowski, 1977; Barclay et al., 1978; Bassili, 1978; Dittrich et al., 1996).

Biological motion research has primarily focused on the perception of human motion because of the significant role that it plays in our social lives. However, our sensitivity to information in motion cues is not restricted to perceiving humans. Humans

Received Feb. 13, 2022; revised Oct. 1, 2022; accepted Oct. 6, 2022.

Author contributions: S.R., L.G.U., and M.V.-P. designed research; S.R. and M.V.-P. performed research; S.R. analyzed data; S.R. wrote the first draft of the paper; S.R., L.G.U., and M.V.-P. edited the paper; S.R. wrote the paper; L.G.U. contributed unpublished reagents/analytic tools.

This work was supported by National Institute of Mental Health Intramural Research Program ZIA-MH-002909. We thank Emalie McMahon for assistance with the optical flow extraction code; Chris Baker for insightful feedback; and Julian De Freitas for inspiring discussions that helped in forming the initial interest in this research area.

The authors declare no competing financial interests.

Correspondence should be addressed to Sophia Robert at srobert@andrew.cmu.edu or Maryam Vaziri-Pashkam at maryam.vaziri-pashkam@nih.gov.

https://doi.org/10.1523/JNEUROSCI.0371-22.2022

can infer animacy and complex social relations from the movements of basic geometric shapes (Heider and Simmel, 1944; Scholl and Gao, 2013; Schultz and Bülthoff, 2013); and can recognize animal categories, such as chickens, dogs, horses, and cats in PLDs (Mitkin and Pavlova, 1990; Mather and West, 1993; Pinto, 1994; Pavlova et al., 2001; Pinto and Shiffrar, 2009).

Investigations of the neural underpinnings of object categorization from motion have identified multiple regions sensitive to lower- and higher-order motion information. The human middle temporal area (hMT^+/V5) has been identified as a motion selective region (Zeki et al., 1991; Watson et al., 1993; Tootell et al., 1995). The posterior superior temporal sulcus has been shown to track animacy signals in motion cues from simple shapes and to process dynamic movements of human faces and bodies (Pavlova et al., 2004; Hirai and Hiraki, 2006; Pitcher et al., 2011; Schultz and Bülthoff, 2013). Neuropsychological studies have suggested that parietal regions are involved in the integration of motion and form information during structure-from-motion identification tasks (Schenk and Zihl, 1997).

Despite extensive research into neural substrates of human motion processing (Giese, 2013), there have been comparatively few studies that have investigated how the motion of nonhuman object categories is processed in the brain. Previous studies suggest preferential processing of human motion over that of one or two other classes (e.g., mammals or tools) in regions in lateral occipito-temporal cortex, including the posterior superior temporal sulcus (Papeo et al., 2017), hMT^+ (Kaiser et al., 2012), and fusiform gyrus (Grossman and Blake, 2002), along with the inferior parietal lobe, inferior frontal gyrus (Saygin et al., 2004), the posterior and anterior cingulate cortices, and the amygdala (Bonda et al., 1996; Ptito et al., 2003).

The limited neuroimaging studies that have directly compared motion- and image-derived object representations have focused on human (or monkey) faces and bodies (Pitcher et al., 2011; Furl et al., 2012; Hafri et al., 2017) or have only compared humans with tools (Beauchamp et al., 2003). Furthermore, these studies (except Beauchamp et al., 2003), have used videos containing both static and dynamic cues as their dynamic condition and therefore could not carefully separate the contributions of motion- and image-information. Thus, a systematic comparison of several object category representations driven by isolated motion and static cues has yet to be undertaken.

Here, we devised a novel method to generate stimuli that only contained motion cues by extracting motion signals from videos of objects and simulating object movements using flow fields of moving dots. Borrowing the term from random dot kinematogram stimuli used to study motion perception, we named our stimuli "object kinematograms" (Cavanagh and Mather, 1989). We first demonstrated that humans can recognize a wide variety of animate and inanimate objects in our dynamic stimuli. We then used these stimuli, along with static images, in an fMRI study to compare object category representations derived from dynamic and static cues in occipito-temporal and parietal regions across visual cortex.

Materials and Methods

Stimuli

Object kinematogram creation pipeline

The process of generating the object kinematograms and their corresponding static images is depicted in Figure 1*A*. Eight categories were selected to sample a wide range of animate and inanimate object categories: human, nonhuman mammal, bird, reptile, vehicle, tool, pendulum/swing, and ball. We searched for videos of objects performing a wide range of movements. Video clips were downloaded from various sources on the Internet or shot with in-house equipment in accordance with the following criteria: (1) contained a single moving object, (2) contained the object in frame without occlusion, (3) shot without camera movement (no zooming, panning, tracking), (4) contained no movement in the background, and (5) lasted at least 3 s.

We used in-house MATLAB code, the Psychtoolbox extension, and in-house Python code to generate moving dot patterns that followed the movement of the objects in the videos. To do this, first, all videos were trimmed to 3 s, cropped with a 3:2 x/y aspect ratio to center the object, and resized to 720×460 pixel resolution. Videos with 30 frames per second were then upsampled so that all videos had a frame rate of 60 fps. The local, frame-by-frame motion of the objects in each video in x and y directions was then extracted using the Farnebäck optical flow algorithm (Farnebäck, 2003).

Next, object movements extracted from the full videos were projected on moving dot patterns. To create the moving dot stimuli, 2500 white dots (2 pixel diameter) were randomly initialized on a gray background (720×460 pixels). Dots that fell within pixels with nonzero motion vector values were moved in the direction and magnitude specified by the extracted motion matrix in the next frame. The lifetime (number of contiguous frames of movement) of any dot was randomly sampled from a uniform distribution between 1 and 17 frames. The lifetime value decreased on every frame. If the lifetime of a dot reached 0 or if they reached the boundaries of the frame, they were reinitialized to a random position with a lifetime of 17 frames.

The number of dots for a given frame and their lifetime was set to mitigate the formation of dot clusters that could induce perception of an edge in individual frames of the video. Individual frames of the videos were qualitatively examined to see whether they induced a perception of any kind of edge, including those related to the object form as well as spurious edges (see example in Extended Data Fig. 1-1). Videos that produced such artifacts were removed from the stimulus set. For the fMRI experiment, these moving dot videos were rendered live for each trial so that the dot initializations were always random.

Object kinematogram validation experiment

To ensure that the stimuli contained clear category information, we conducted an online experiment; 430 participants (223 women, aged 18-65 years) were recruited on Amazon Mechanical Turk to perform an object categorization task on the dynamic stimuli. Participants each performed either 10 or 11 trials. For each trial, participants were asked three questions about the object in a looped video: (1) whether the object in the video was of an animal or non-animal, (2) which of 8 listed categories the object belonged to, and (3) whether they could label the object. If subjects responded "yes" for the third question, they were required to type the label in a response text box. Each of the three questions contained an "I don't know" option. Subjects had to answer all three questions to complete each trial.

Overall, subjects categorized objects based on their motion in the moving dot stimuli with an average accuracy of 76% (202 total videos). The three animate (human, mammal, reptile) and three inanimate (tool, ball, pendulum/swing) categories with the highest accuracy were used for the fMRI experiment. For each category, the six videos with the highest accuracy were selected (mean accuracy = 96%).

The overall "motion energy" of each video was calculated by averaging the motion vectors across all pixels in all frames. The average motion energies for the 6 videos in each category were entered into pairwise two-sample heteroscedastic t test comparisons to ensure that there were no significant differences between categories (all uncorrected p values > 0.05; for mean and SDs per category, see Table 1).

After the object kinematogram stimulus set was finalized, the static image stimulus set was generated by randomly selecting three frames of the full form video from which the moving dot stimulus was created. The frame with the object in clearest view was selected and further processed to extract the object from the frame. For the fMRI experiment, the isolated object was pasted onto a background of 2500 randomly initialized white dots on a gray background to mimic a frame of the dynamic moving dot stimuli (Fig. 1*A*).

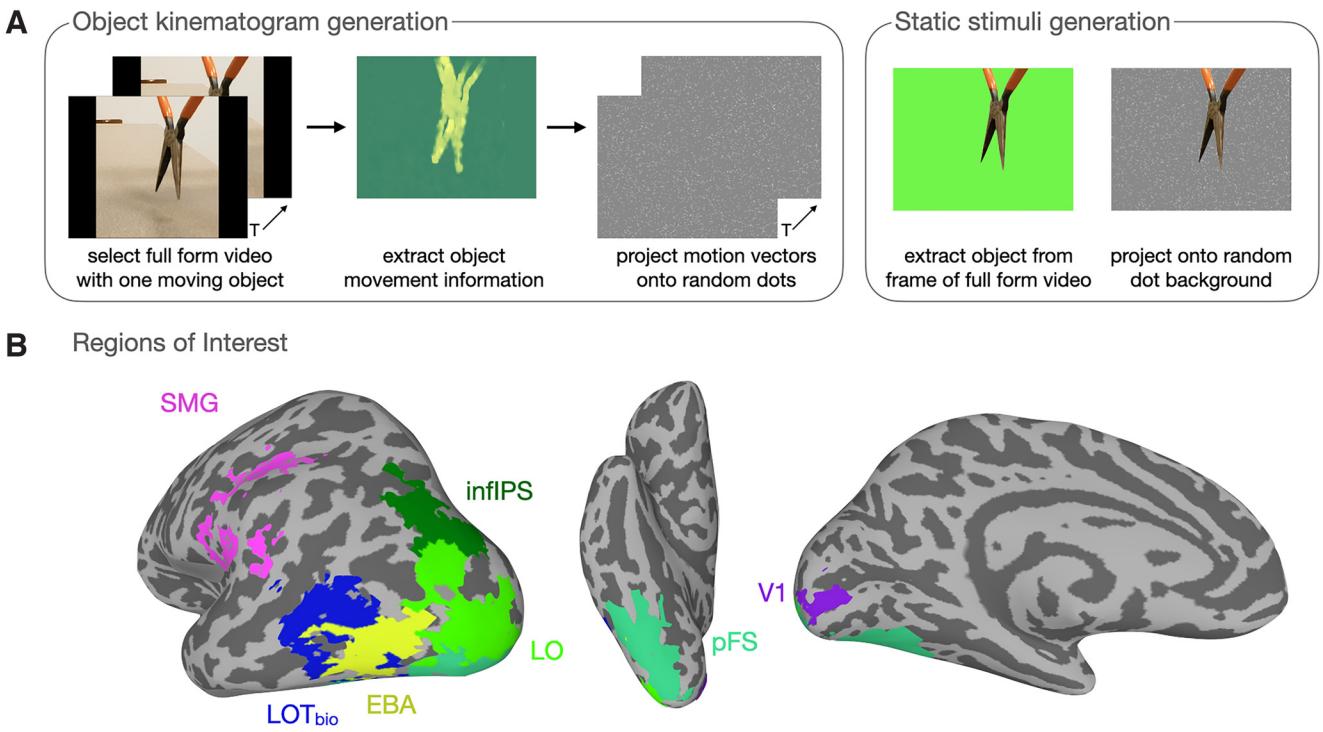

Figure 1. Schematic depicting the stimulus generation process and regions of interest (ROIs) of a single example subject generated by the group-constrained single-subject method. ***A***, General pipeline for generating the object kinematogram stimuli (left) and static image stimuli (right). "T" indicates the time between two frames of the video as the extracted motion information is calculated with pairs of frames. Kinematograms that induced perception of form in a static frame were eliminated (see Extended Data Fig. 1-1). ***B***, ROIs for one example subject. Pink represents the supramarginal area (SMG). Dark green represents the inflPS. Light green represents the LO. Yellow represents the EBA. Dark blue represents the biological motion-related lateral occipito-temporal area (LOT-biomotion). Teal represents the pFS. Purple represents primary visual cortex (V1). As an STS region could not be reliably localized using the group-constrained single-subject method, we also defined LOT and STS regions (LOT-atlas and STS-atlas) from a probabilistic atlas (shown in Extended Data Fig. 1-2). The overlap of the SMG region with meta-analytic activation maps for "action observation" and "tools" from Neurosynth (Yarkoni et al., 2011) is shown in Extended Data Figure 1-3.

Table 1. Motion energies of dynamic stimuli per object category

Motion energy	Humans	Mammals	Reptiles	Tools	Pendulums	Balls
Mean	0.1298	0.2191	0.1016	0.3441	0.1303	0.1327
SD	0.0364	0.1836	0.0520	0.2536	0.1403	0.0506

fMRI experiment

Participants

Fifteen healthy human subjects (6 female, age range 19-42 years) with normal or corrected to normal vision were recruited for the fMRI experiment. Based on effect sizes from other studies using multivariate decoding (Cohen's d values > 7.3 for cross-classification of fMRI responses to images across position and size in ROIs in lateral and ventral occipito-temporal cortex; Vaziri-Pashkam and Xu, 2019), we conducted a power analysis and concluded that a sample size of 15 would be sufficient to detect an effect size of ≥ 0.9 with a power of 0.8. Participants were brought in for a 2 h fMRI session that included the main experiment and three localizer tasks. Before entering the scanner, all participants practiced the tasks for the main experiment and localizer runs and underwent a short behavioral task to familiarize themselves with the stimuli. All subjects provided informed consent and received compensation for their participation. The experiments were approved by the National Institutes of Health ethics committee.

Training session

The independent norming study performed with mTurk demonstrated that people can recognize the objects in these stimuli with high accuracy after minimal instruction. However, to avoid introducing noise because of intersubject variability in recognition of the dynamic stimuli or potentially slower recognition of the dynamic stimuli in the first runs of the session, participants performed a training session before entering the scanner. During the training session, they familiarized themselves with the 36 dynamic stimuli and were subsequently tested to ensure accurate recognition. Each video was shown on loop until subjects could verbally report which of the 6 categories the object belonged to. If the subject categorized the object correctly, the experimenter advanced to the next stimulus; incorrect categorizations were verbally corrected by the experimenter. After all stimuli had been verbally categorized, subjects underwent a testing session. In each trial, a random video was shown once without looping, followed by a gray screen with six category labels placed in a circle around the center of the screen. Subjects were instructed to categorize the object in the video by clicking on the corresponding category label. No feedback was provided during the testing session. If a subject performed above 90% accuracy, they continued on to the fMRI experiment. The training and testing session took no longer than 15 min. Subjects required little to no correction during the training session and performed with an average of 99% accuracy in the test session on the first iteration ($n = 13$, data for 2 subjects were lost because of technical problems).

MRI methods

MRI data were collected from a Siemens MAGNETOM Prisma scanner at 3 Tesla equipped with a 32-channel head coil. Subjects viewed the display on a BOLDscreen 32 LCD (Cambridge Research Systems, 60 Hz refresh rate, 1600 × 900 resolution, at an estimated distance of 187 cm) through a mirror mounted on the head coil. The stimuli were presented using a Dell laptop with MATLAB and Psychtoolbox extensions (Brainard, 1997; Kleiner et al., 2007).

For each participant, a high-resolution (1.0 × 1.0 × 1.0 mm) T1-weighted anatomic scan was obtained for surface reconstruction. All functional scans were collected with a T2*-weighted single-shot, multiple gradient-EPI sequence (Kundu et al., 2012) with a multiband acceleration factor of 2 slices/pulse; 50 slices (3 mm thick, 3 × 3 mm^2 in-plane resolution) were collected to cover the whole brain (TR 2 s, TE = 12 ms, 28.28 ms, 44.56 ms, flip angle = 70°, FOV = 216 mm).

Experimental design

Main experiment. The main task of the experiment included 6 categories (human, mammal, reptile, tool, pendulum/swing, and ball) and 2 stimulus conditions: dynamic (object kinematograms) and static (object images pasted on dot background). Both dynamic and static stimuli were presented at the same size and location (subtending 9.6° × 4.8° visual angle). We used a block design to present alternating blocks of dynamic and static stimuli while also alternating between animate and inanimate blocks. The order of the 6 categories and the two formats were counterbalanced within and across runs. Four different counterbalancing designs were created, and each subject was randomly assigned one of the designs.

Each run contained 12 condition blocks, one for each condition (2 formats × 6 categories), began with an initial fixation block of 8 s, and ended with a final fixation of 12 s. Each condition block began with an 8 s fixation period in which a red fixation dot (5 pixels in radius) was shown on a gray background. The fixation period was then followed by the stimulus presentation period in which 4 stimuli were presented from the same condition, each for 2.8 s followed by a 200 ms interstimulus interval, resulting in 12 s of stimulus presentation. The duration of each condition block was 20 s (8 s fixation and 12 s stimulus presentation). For each run, the 12 condition blocks and the initial and final fixation blocks lasted 252 s (4 min 12 s). Each participant completed 12 runs.

To maintain their attention, subjects were given a one-back repetition detection task in which they were instructed to press a button on an MRI-compatible button box (fORP, Cambridge Research Systems) to indicate the detection of a repeated stimulus within each block. There was one stimulus repetition per block, and the repeated stimulus of each block type was changed across runs. Because there were only three unique trials per block but each condition had 6 unique stimuli, half of the stimuli of each category were shown on odd runs and the other half were shown on the even runs. These blocks were later combined during analysis. Average performance on this task was 94%. To ensure proper fixation, eye movements were monitored using an ASL eye-tracker.

Object localizer task. To localize functional ROIs in the ventral and lateral occipito-temporal cortex, we presented images of objects in 6 conditions: faces, scenes, head-cropped bodies, central objects, peripheral objects (four objects per image), and phase-scrambled objects in a block design paradigm. Subjects were instructed to fixate while 20 images were presented in each block for 750 ms with a 50 ms fixation screen in between. Each block lasted 16 s and was repeated 4 times per condition. Each run started with a 12 s fixation period. Additional 8 s fixation periods were presented after every 5 blocks. Total run duration was 436 s (7 min 16 s). Subjects performed a motion detection task. During each block, a random image would jitter by rapidly shifting 4 pixels back and forth horizontally from the center of the screen. Subjects indicated detection of motion with a button press. Each participant completed 1 or 2 runs of this task.

Motion localizer task. To localize functional ROIs related to the perception of biological and nonbiological motion, we presented blocks of PLD videos of humans performing various actions in four conditions: (1) biological motion: normal PLD video of walking, riding a bicycle, etc.); (2) random motion: the points in the PLD were spatially scrambled in each frame; (3) translation: randomly positioned dots translated across the screen in a random direction with their speed set to the average speed of the movement from the PLD videos; and (4) static: a random frozen frame of the PLD was shown as an image. There were 8 exemplars per condition, each presented for 1.5 s followed by a 500 ms interstimulus fixation period. Each block lasted 16 s and was presented 4 times per condition. Each run began with a 6 s fixation period, and 8 s fixation periods were interspersed between each block, making the total run duration 422.7 s (7 min 3 s). Subjects performed a one-back repetition detection task, in which they indicated detection of a repeated stimulus during each block by pressing a button. Each subject completed 1 or 2 runs of this task.

Topographic mapping. Topographic visual region V1 was mapped using 16 s blocks of a vertical or horizontal polar angle wedge with an arc of 60° flashing black and white checkerboards at 6 Hz. During the stimulus blocks, subjects fixated on a red fixation dot (5 pixel radius) and detected a dimming on the wedge, that occurred randomly either at the inner, middle, or outer ring of the wedge at four random times within the 16 s block. There was a 16 s fixation period after each block, and each run began with a 16 s period of fixation. Each run lasted 272 s (4 min and 40 s), and subjects completed 1 or 2 runs of this task.

Data analysis

fMRI data were analyzed using AFNI (Cox, 1996) and in-house MATLAB codes. The data were preprocessed by removing the first 2 TRs of each run, motion correction, slice timing correction, smoothing with 5 mm FWHM, and intensity normalization. The EPI scans were registered to the anatomic volume using the first volume of the first run and the default algorithm from AFNI. The three echoes were optimally combined using a weighted average (Posse et al., 1999; Kundu et al., 2012). TRs with motion exceeding 0.3 mm and outliers, defined as time points where >10% of voxels in the brain mask deviated substantially from the time series trend (using the default settings of 3dToutcount), were excluded from further analysis. A GLM analysis with 12 factors (2 stimulus conditions × 6 categories) was used to extract β and t values for each condition in each voxel. Movement parameters with 6 degrees of freedom were used as an external regressor. To account for the effect of residual autocorrelation on statistical estimates, we applied a generalized least squares time series fit with restricted maximum likelihood estimation of the temporal auto-correlation structure in each voxel. Beta values were calculated across all runs for the univariate analysis, and t values were calculated per-run for the multivariate analysis.

ROI definition: group-constrained subject-specific method

The retinotopy task was used to identify primary visual cortex (V1), to serve as a control region. The object category localizer task allowed us to localize canonical ROIs in occipito-temporal cortex known to process object category information, including lateral occipital cortex (LO), posterior fusiform sulcus (pFS), and the extrastriate body area (EBA). We also included inferior intraparietal sulcus (infIPS), a dorsal region that has been implicated in object individuation (Xu and Chun, 2009) and has been shown to contain abstract object category representations (Vaziri-Pashkam and Xu, 2017; Vaziri-Pashkam et al., 2019). Using the biological motion localizer, we identified regions selective for biological motion over translation in LOT-biomotion. Because of the variability of the STS biological motion region in our subjects, we were not able to localize this region using our ROI selection method described below. For completeness, we have included an STS region based on a probabilistic atlas of point-light-display responsive regions (Engell and McCarthy, 2013) in our Extended Data (for an overlap map of this region with our LOT-biomotion region, see Extended Data Fig. 1-2; for the univariate and multivariate results, see Extended Data Fig. 2-1). Finally, as we did not have a localizer for a tool-selective region, we included an atlas defined supramarginal region (SMG) because it overlaps with the parietal tool-selective region (Peeters et al., 2009, 2013). This region has also been implicated in action observation (Rizzolatti and Craighero, 2004; Caspers et al., 2010). An overlap map of our SMG ROI with meta-analytic maps generated from studies investigating terms related to "action observation" and "tools" from Neurosynth (Yarkoni et al., 2011) can be found in Extended Data Figure 1-3.

Figure 1*B* shows the final ROIs for one example subject. We used a systematic, unbiased method for creating individualized ROIs constrained by group responses to our localizer experiments, basing our approach on a method of ROI definition developed by Kanwisher and Fedorenko (described in Pitcher et al., 2011). First, t values were extracted from GLMs of individual activation maps from the localizer experiments. All subjects' statistical activation maps ($N = 15$) were converted to Talairach space. For each subject, the individual localizer contrast maps were thresholded at $p < 0.0001$. Group overlap proportion maps were then created for each contrast. Second, we thresholded the group proportion maps for each contrast separately to counteract contrast- or localizer-specific differences in spatial variability or overall activation. The thresholds for specific contrast maps were as follows: For the object localizer experiment, the thresholds were $N \geq 0.7$ for objects versus scrambled (LO and pFS), $N \geq 0.5$ for bodies versus objects (EBA),

and $N \geq 0.25$ for peripheral objects versus scrambled (infIPS). For the biological motion experiment, the threshold for biological motion versus translation was $N \geq 0.5$ (LOT-biomotion). For the retinotopy experiment, positive and negative maps were created separately and thresholded at $N \geq 0.5$ (V1). We then used a Gaussian blur of 1 mm FWHM. The blurred maps were then clustered using the nearest neighbors method and a minimum cluster size of 20 voxels. For V1, positive and negative maps were clustered separately and then combined with a step function. Two steps were required to finalize the group-constrained ROIs. Anatomical landmarks were used to separate pFS from LO, and LO from infIPS. V1 was separated from V2 using a hand-drawn region based on the group map. All ROIs were then selected to have no overlapping voxels.

The final nonoverlapping group-constrained ROIs were made subject specific by creating masks based on the individual subject's activity during the localizer experiments (localizer contrast threshold: $p < 0.05$). For example, for each subject's EBA, the group-constrained EBA was masked by the subject's response to bodies > objects with a threshold of $p < 0.05$. If this process did not yield an ROI with at least 100 voxels across the two hemispheres, the ROI was instead created with a mask made from the mean response during the main experiment (task vs fixation, $p < 0.0001$ uncorrected).

The supramarginal (SMG) ROI was anatomically defined using a Freesurfer parcellation (Desikan et al., 2006). To make the subject specific supramarginal ROIs, individual masks were made from the mean response during the main experiment (task vs fixation, $p < 0.0001$ uncorrected) and intersected with the template SMG region.

Univariate analysis

To calculate the average fMRI response per condition for each ROI, using a GLM analysis, whole-brain β value maps were extracted for each of the 12 conditions and masked with a task > fixation threshold of $p < 0.0001$ for each subject. The group-constrained subject-specific ROIs were intersected with these maps, resulting in a β value response per voxel in each ROI for all 12 conditions in each subject. Because of previous work demonstrating animacy as an organizing principle of object category information within occipito-temporal cortex (Beauchamp et al., 2003; Kriegeskorte et al., 2008; Konkle and Caramazza, 2013), for each ROI, we were interested in differences between univariate responses to the animate and inanimate categories as well as stimulus format. The average responses for four conditions were calculated for each ROI: dynamic animate, dynamic inanimate, static animate, and static inanimate. The animacy preference was calculated as the difference between the animate and inanimate conditions, separately for the static and dynamic stimulus formats. A two-way ANOVA was conducted for each ROI, evaluating the main effects of stimulus format and animacy and their interaction. Two-sided one-sample t tests were conducted to determine whether the animacy preference in each ROI, and each format was significantly different from 0. All t tests were corrected for multiple comparisons with false discovery rate correction (Benjamini and Hochberg, 1995) across ROIs. For ANOVAs, effect sizes were calculated with generalized η squared (η^2_G); for the t tests, Cohen's d was used.

Multivariate pattern analysis

We performed multivariate pattern analyses to investigate whether object category information was present in the fMRI responses to the dynamic and static stimuli. We extracted t values in each voxel for every condition in each run using a GLM analysis. To perform pairwise object category decoding, we used a linear support vector machine classifier (SVM) (Chang and Lin, 2011) with feature selection and a fixed regularization parameter set to the default value. The SVM was trained using leave-one-run-out cross validation on data that was normalized with z scoring to avoid magnitude differences between conditions. Using t tests, we calculated the top 100 most informative voxels per ROI (Mitchell et al., 2004) to equate the number of voxels analyzed per ROI and facilitate comparisons between them. This feature selection was performed separately for each iteration of training. Results did not qualitatively change when the analysis was performed without feature selection.

We trained and tested the linear SVM in two conditions: (1) within-classification, in which the SVM was trained and tested on the same stimulus format; and (2) cross-classification, in which SVM was trained in one stimulus format and tested on the other format. The classification was performed on all unique pairs of object categories to obtain classification accuracy matrices. The off-diagonal values of the matrices were averaged to produce two within-format (dynamic and static) and two cross-format (train dynamic test static, train static test dynamic) average object category decoding accuracies per subject. The two cross-format values were then averaged to obtain one cross-classification accuracy. Two-sided one-sample and paired t tests were conducted to determine, respectively: (1) if the decoding accuracy in each ROI and each format was significantly different from chance (0.5), and (2) if the decoding accuracy was significantly different across stimulus formats within each ROI. All p values listed from t tests were corrected for multiple comparisons with false discovery rate correction across ROIs (Benjamini and Hochberg, 1995). For ANOVAs, effect sizes were calculated with generalized η^2_G. For the one-sample and paired t tests, Cohen's d was used.

To ensure that our ROIs provided sufficient coverage of regions that process static and dynamic visual input, we conducted a whole-brain searchlight decoding analysis for object category responses to both stimulus formats. Searchlight maps were generated for each subject using the Gaussian Naive Bayes searchlight classifier from the Searchmight Toolbox version Darwin i386.0.2.5 in MATLAB) (Pereira and Botvinick, 2011). The Gaussian Naive Bayes classifier was chosen over a linear SVM to increase computational speed. Group maps were generated by conducting a two-sided t test of the subjects' accuracy maps against 0.5 using AFNI's 3dttest++, separately for the within dynamic, within static, and across-format decoding. The across-format maps were an average of the decoding accuracies for training on dynamic and testing on static and vice versa. The group results for each map were thresholded with a q value at 0.005 for the within-format maps and 0.05 for the across-format map.

Multidimensional scaling of fMRI responses

To visualize how stimulus format and object category impact the responses in our ROIs, we quantified the similarities between the patterns of fMRI responses to the 12 conditions in each ROI by calculating all pairwise Euclidean distances. The individual subject Euclidean distances per ROI were averaged across subjects to create group Euclidean distances, which will be referred to as the fMRI-Euclidean matrix. We then visualized these similarities by applying classical multidimensional scaling (Shepard, 1980) on the fMRI-Euclidean matrix and plotting the first two dimensions for each ROI.

We measured the reliability of the fMRI-Euclidean matrix by performing a leave-one-subject-out (LOSO) analysis wherein an individual subject's matrix was correlated with the remaining subjects' average matrix (Op de Beeck et al., 2008; Nili et al., 2014). Pearson's correlations of each subject were averaged across iterations to produce a final reliability score and its standard error. The reliabilities of the dynamic and static fMRI-Euclidean matrices were evaluated separately.

Object similarity behavioral experiment

A total of 353 participants (32% female among the 85% who responded to the demographic survey) were recruited on Amazon Mechanical Turk to perform an object similarity task on the dynamic or static stimuli. All participants were located in the United States.

For each trial, participants were presented with three stimuli on a gray screen and were instructed to select the odd-one-out stimulus (the stimulus that was most distinct among the three) by clicking on it. Dynamic and static stimuli were tested separately. Participants performed blocks of 15 trials to complete the task and were permitted to perform more than one block. To ensure data quality, trials with reaction times < 0.6 s and 1.2 s and > 10 s or 20 s were removed for the image and video tasks, respectively. These cutoffs were decided based on the distributions of reaction times. If 5 or more trials in a block were eliminated, the entire block (or HIT in mTurk terminology) was removed. The eliminated blocks were resubmitted to mTurk to ensure that we had at least 2 repetitions for each unique triplet allowing for 68 trials for each pair of stimuli.

To build a dissimilarity matrix based on the odd-one-out image and video tasks, a response matrix of the pairwise dissimilarity judgments was constructed for each task by treating each triplet as three object pairs and assigning 1's to dissimilar pairs (i.e., the two pairs that included the selected odd object) and a 0 to the similar pair (i.e., the pair that did not include the selected odd object). We also constructed a count matrix to determine how many times each pair was shown together in a triplet. By dividing the response matrix by the count matrix, we obtained a dissimilarity matrix with values ranging from 0 to 1 with higher values denoting higher dissimilarity. To produce a category level behavioral dissimilarity matrix, we took the off-diagonal upper triangle of the 36×36 matrix and averaged the item distances that belonged to the same category, resulting in a 6×6 matrix, which will be referred to as the behavioral-dissimilarity matrix. The diagonal was nonzero because of nonzero distances between exemplars within each category. Only the off-diagonal of this matrix was used in further analyses.

To gauge the stability of the behavioral-dissimilarity matrix, we performed a split-half reliability analysis. Because each subject only saw a small set of all possible triplets, instead of splitting the data by subject, we split based on repeats of stimulus pairs (three pairs per triplet) into two groups. The binary similarity values for all pairs were correlated across the two groups to produce a measure of reliability of the similarity judgments.

Multidimensional scaling and hierarchical clustering of object similarity responses

We visualized the structure of the object similarity judgments from the odd-one-out tasks at the category level using classical multidimensional scaling on the behavioral-dissimilarity matrices of the dynamic and static stimuli separately (Shepard, 1980). The two behavioral-dissimilarity matrices were also correlated to quantify their degree of similarity. To investigate the structure of the object similarity judgments at the exemplar level, we used a hierarchical or agglomerative clustering algorithm available in the Python package *SciPy* (Virtanen et al., 2020) on the dynamic and static behavioral-dissimilarity matrices separately. For visualization purposes, images of the individual exemplars, which were adapted from the static stimuli used in the experiment, were included under the resultant dendrograms for both static and dynamic conditions (note that dynamic stimuli are not recognizable in static frames).

Brain-behavior correlation

To determine the relationship between the multivariate information for the 6 categories in each ROI (fMRI-Euclidean matrix) and behavioral assessments of the category similarity (behavioral-dissimilarity matrix), we correlated the two measures. For each subject, the off-diagonal of the fMRI-Euclidean matrix was correlated with the off-diagonal of the behavioral-dissimilarity matrix using Pearson's linear correlation coefficient, separately for the dynamic and static experiments. The correlations were then averaged across subjects. To take into account the combined noise from the two measures, the noise ceiling was calculated for each ROI as $\sqrt{R1 \times R2}$, or the square root of the product of the reliabilities of the fMRI-Euclidean matrix (R1, LOSO) and the behavioral-dissimilarity matrix (R2, split-half), as used in previous studies (Nunnally, 1970; Op de Beeck et al., 2008; Vul et al., 2009).

Brain-optic flow correlation

To ensure that optic flow information from the 6 object categories was not predictive of the multivariate fMRI responses in any of the ROIs, we performed a control analysis. We first calculated the Euclidean distances between the dynamic stimulus information of each category by vectorizing the 4-dimensional stimuli (x-coordinates, y-coordinates, x- and y-magnitudes of optic flow, and time: $460 \times 720 \times 2 \times 180 = 119{,}232{,}000$ element vector) and averaging the distances between stimuli of the same category, creating the optic flow-Euclidean matrix. We then correlated the optic flow-Euclidean matrix with the dynamic fMRI-Euclidean matrix of each ROI for each subject. Two sided one-sample t tests were used to determine whether any positive correlations were significantly different from zero.

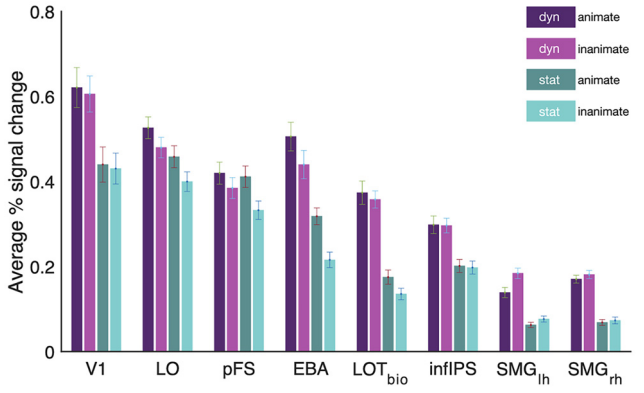

Figure 2. Univariate fMRI responses to dynamic animate (purple) and inanimate (pink) stimuli and static animate (dark green) and inanimate (teal) stimuli for each ROI. Results did not qualitatively differ when removing the human and tool categories from the analysis. Also, see Results from the LOT-atlas and STS-atlas regions defined by a probabilistic atlas in Extended Data Figure 2-1*A*. Error bars indicate standard error of the mean.

Code accessibility

The stimuli and custom codes used in this study can be accessed through the OSF website (https://osf.io/45b8y/). The scripts for stimulus presentation and the object kinematograms used in the study are available under "Manuscript Scripts" and "Manuscript Stimuli," respectively. The scripts used to generate the object kinematograms were included in the "Object kinematogram generation code" at https://osf.io/eqd87/, along with instructions and recommendations for generating optimal stimuli. Analysis code and data will be made available on request.

Results

Effect of stimulus format on univariate animacy preference

We first looked at the mean amplitude of responses to the two superordinate object categories (animate/inanimate) in the two stimulus formats (static/dynamic). We extracted individual subjects' β values from the GLM analysis and averaged the response for the 3 animate and the 3 inanimate categories within each image format to get 4 values per subject. Figure 2 shows the pooled results of this analysis across subjects. A two-way ANOVA with stimulus format and animacy as factors showed a significant main effect of stimulus format in all ROIs (F values > 5.63, p values ≤ 0.03, η^2_G values > 0.03) with higher response amplitudes in the dynamic compared with the static condition. A main effect of animacy was also found in LO, pFS, EBA, LOT-biomotion, and left SMG (F values > 8.34, p values ≤ 0.02, η^2_Gs > 0.03), but not in V1, infIPS, or right SMG (F values < 2.43, p values > 0.63, η^2_G values < 0.0004).

We subtracted inanimate responses from animate responses to produce a measure of animacy preference within each stimulus format (Fig. 3). As a *post hoc* comparison of the main effect of animacy, we conducted two-sided one-sample t tests of animacy preference against 0 (Fig. 3). For both stimulus formats, LO, pFS, and EBA showed a preference for animate categories (dynamic: t values > 2.84, p values < 0.03, Cohen's d values > 0.76, static: t values > 6.76, p values < 0.001, Cohen's d values > 1.81), while left SMG preferred inanimate categories (dynamic: $t_{(14)} = 4.58$, $p = 0.002$, Cohen's $d = 1.22$). LOT-biomotion had a significant preference for animate categories in the static ($t_{(14)} = 4.13$, $p = 0.002$, Cohen's $d = 1.11$) but not in the dynamic condition ($t_{(14)} = 1.28$, $p = 0.30$, Cohen's $d = 0.34$).

Further, there was a significant interaction between animacy and stimulus format for pFS and left SMG (F values > 8.72, p values < 0.04, η^2_Gs > 0.01), and not for the rest of the regions (F

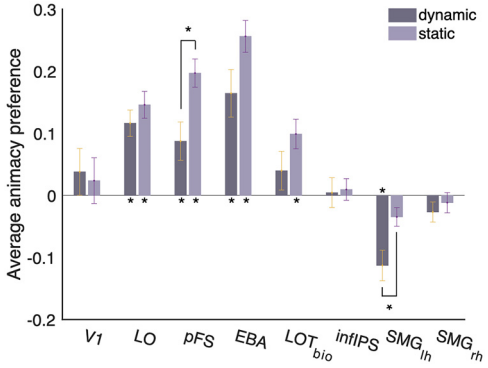

Figure 3. Univariate fMRI response preference for animate compared with inanimate object categories in dynamic (dark gray) and static (light gray) formats for each ROI. Also, see Results from the LOT-atlas and STS-atlas regions defined by a probabilistic atlas in Extended Data Figure 2-1*B*. *$p < 0.05$. Error bars indicate standard error of the mean.

values ≤ 6.35, p values > 0.06, η_G^2s < 0.008). pFS, a ventral region known to be involved in object recognition from static images, showed a stronger preference for animate object stimuli in the static compared with the dynamic condition ($F_{(14)} = 10.38$, $p = 0.04$, $\eta_G^2 = 0.01$), while left SMG, a parietal region thought to be involved in tool processing and action observation, had a stronger preference for inanimate object stimuli in the dynamic compared with the static condition ($F_{(14)} = 8.72$, $p = 0.04$, $\eta_G^2 = 0.04$). These significant interactions between stimulus format and animacy preference suggest that the category preference responses in pFS and left SMG are modulated by the format through which the category information is provided. The most ventral region, pFS, is more sensitive to static form presentations of animate objects; and the most dorsal lateral region, left SMG, is more sensitive to dynamic motion information about inanimate objects.

Effect of stimulus format on multivariate object category representations

We next examined the multivariate patterns of each of our ROIs to further explore how object category information is represented in the brain when sourced from dynamic movements and static images. We first sought to test whether each of our regions contained information about the 6 object categories within each stimulus format. To do this, we calculated the average pairwise classification accuracy for the 6 object categories for the static and dynamic conditions using a linear SVM classifier (Chang and Lin, 2011). Figure 4*A* shows the pooled results of this analysis across subjects. Two-sided one-sample t tests revealed that the object categories were decoded significantly above chance in both dynamic and static formats in all regions but V1 (dynamic: t values > 7.04, $p < 0.001$, Cohen's d values > 1.82; static: t values > 2.73, p values < 0.02, Cohen's d values > 0.71). In V1, significant decoding was only found in the static stimulus condition (static: $t_{(14)} = 8.31$, $p < 0.001$, Cohen's $d = 2.15$; dynamic: $t_{(14)} = 2.05$, $p = 0.06$, Cohen's $d = 0.53$). In all regions but infIPS, there were significant differences between the decoding accuracies across stimulus format (infIPS: $t_{(14)} = 0.59$, $p = 0.57$, Cohen's $d = 0.15$). In V1, LO, pFS, and EBA decoding accuracies were higher in the static condition than the dynamic (t values > 2.32, p values < 0.001, Cohen's d values > 0.60); while in LOT-biomotion and bilateral SMG, decoding accuracies were higher in the dynamic condition (t values > 3.24, p values < 0.008, Cohen's d values > 0.84).

To ensure that the significant decoding of object category from dynamic information was because of differences in the responses to object categories and not contingent on differences in optic flow information that were confounded with category in our stimulus set, we performed a control analysis in which we correlated the dynamic stimulus information with the multivariate fMRI responses (see Materials and Methods). No significant positive correlations were observed for any of the ROIs (t values < 2.8, p values > 0.06). Because the 6 exemplars per category were systematically split across runs, we were also able to calculate the decoding accuracy across splits of the data corresponding to the two sets of stimuli within the same object category (Extended Data Fig. 4-1). This analysis mirrored our original findings: we found robust, albeit lower, generalization across the two sets of stimuli within the same category in all ROIs, with the exception of V1 for the dynamic condition and bilateral SMG for the static condition.

We next used a cross-classification method to determine whether abstract responses to object categories, regardless of stimulus format, were present in our ROIs. The SVM classifier was trained in one stimulus format and then tested in the other format. Decoding accuracies when training on static and testing on dynamic and training on dynamic and testing on static were averaged to produce the light gray bars shown in Figure 4*B*. We also calculated the within-classification accuracy for training and testing within stimulus format (dark gray bars in Fig. 4*B*; average of the two bars in Fig. 4*A*). Robust within-format classification accuracy in all regions was observed, even when training and testing across subsets of the stimuli within each category (Extended Data Fig. 4-1). Significant cross-classification was observed in all ROIs (t values > 5.31, p values < 0.0001, Cohen's d values > 1.37), and was significantly lower than within-classification in all ROIs (t values > 5.24, p values < 0.0001, Cohen's d values > 1.35).

These across-format classification results and our control analyses suggest that there is sufficient abstract information about object categories in the multivariate pattern responses to the dynamic and static stimuli to allow for generalization across stimuli and formats in regions across visual cortex. In some of the regions, including LO, pFS, LOTbio, and infIPS, across-format decoding was higher when training on the dynamic and testing on the static condition (see the breakdown of training and testing in each direction in Extended Data Fig. 4-2). This was likely observed because the object kinematograms were better controlled for low-level features than the static stimuli, facilitating across-format decoding, which requires that sufficient information about object category is available in the training set. Furthermore, a whole-brain searchlight analysis of within- and across-format decoding using the Searchmight Toolbox (Pereira and Botvinick, 2011) demonstrated that our ROIs cover the primary loci of object category information derived from dynamic and static visual inputs (Fig. 4*C–E*).

To further visualize the similarity between the fMRI responses to the object categories in the dynamic and static conditions, we calculated the pairwise Euclidean distances between the patterns of responses to the 6 object categories and the 2 stimulus formats in each ROI. We then performed a multidimensional scaling analysis on the resultant dissimilarity matrix and visualized the first two dimensions in each of the ROIs (Fig. 5). In all regions, there was a clear separation between the responses to the dynamic (shown in purple and pink) and static stimuli (shown in green and teal). In addition, the ventro-temporal regions and inferior parietal cortex showed a separation among the individual object

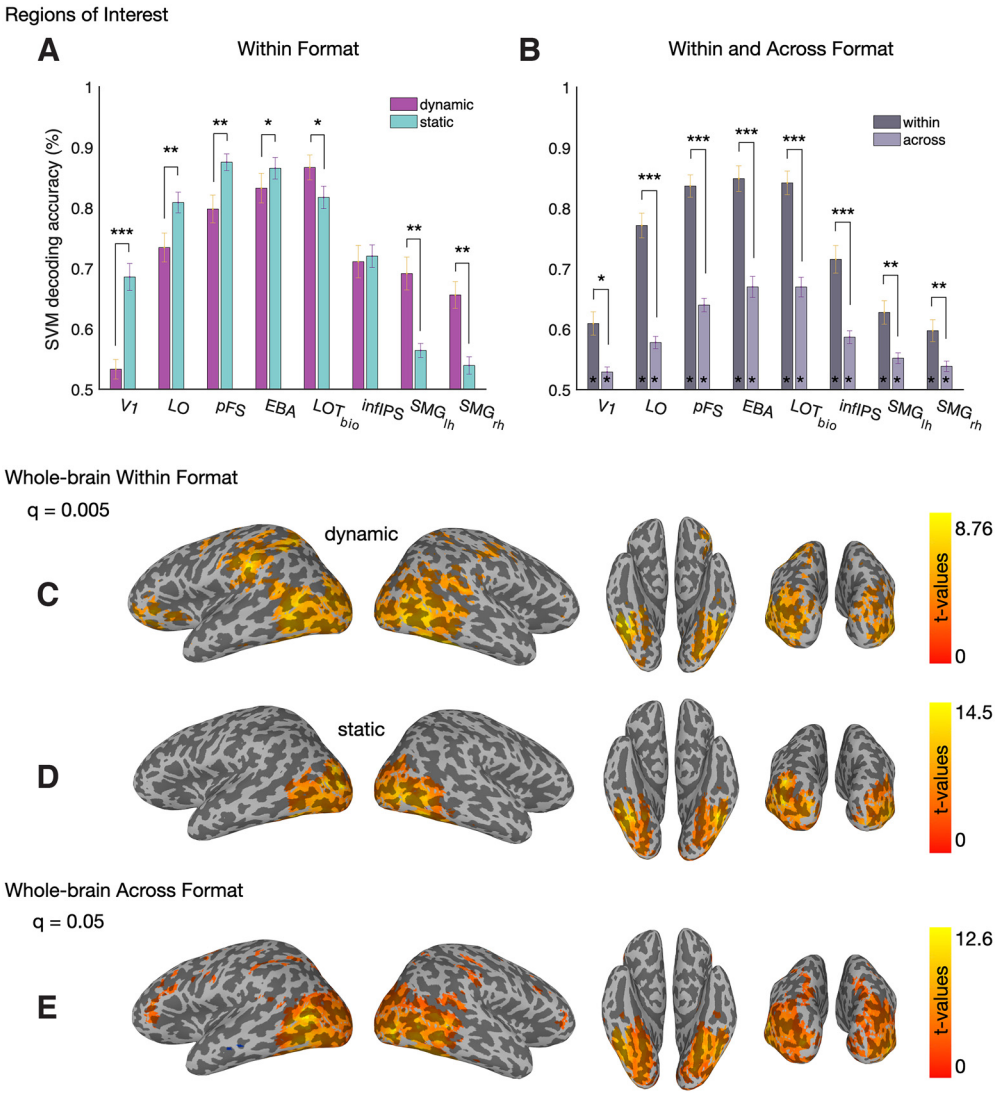

Figure 4. Object category SVM decoding accuracies in each ROI and across the whole brain. ***A***, Average SVM decoding accuracies when training and testing within the dynamic (pink) and static (teal) conditions. All average decoding accuracies were significantly above chance except for the dynamic condition in V1. Above chance decoding was also found when generalizing across subsets of the stimuli within each category in the dynamic and static formats (Extended Data Fig. 4-1). ***B***, Average within (dark) and across (light) format decoding accuracies. Within stimulus format decoding accuracies were produced by averaging the dynamic and static decoding accuracies in ***A***. Cross classification was significantly above chance in all ROIs. Above chance cross classification was observed when training on the dynamic and testing on the static format, and vice versa (Extended Data Fig. 4-2). See also results from the LOT-atlas and STS-atlas regions defined by a probabilistic atlas in Extended Data Fig. 2-1*C*, *D*. Error bars indicate standard error of the mean. Asterisks within the bars represent significance in *t* tests against chance. Asterisks above bars represent paired *t* tests across conditions. $*p < 0.05$. $**p < 0.001$. $***p < 0.0001$. ***C–E***, Whole-brain group *t* value maps for decoding (***C***) within dynamic, (***D***) within static, and (***E***) across stimulus format. Whole-brain searchlight analysis showed results consistent with our ROI analysis. Color bars represent the range of *t* values for the group comparison against 0.5. Within-format maps were thresholded at $q = 0.005$, and the across-format map was thresholded at $q = 0.05$.

categories. The nearly parallel lines connecting the dynamic and static conditions of the same category indicate that categories with responses that were similar to each other in one condition were also similar to each other in the other condition and are in line with the results of the cross-classification analysis performed earlier. In bilateral supramarginal areas, this object category separation was evident for the dynamic stimulus responses, but the static stimulus responses remained clustered together. In V1, while there was a separation between dynamic and static conditions, the arrangement of categories does not appear to be consistent across conditions.

Odd-one-out behavioral experiment

To investigate how the responses of each ROI to the 6 object categories in each format relate to the behavioral measure of similarity, we performed two behavioral experiments on Amazon Mechanical Turk. We showed participants three objects (either in static format or in dynamic format) and asked them to judge the similarity between the three objects and pick the odd-one-out. We calculated two dissimilarity matrices based on the responses: one for the static stimuli and one for the dynamic stimuli (see Materials and Methods). We then averaged the individual object distances from each category to obtain dissimilarity scores between the 6 object categories for the 2 stimulus formats (Fig. 6*A*). The reliability of these similarity judgments was evaluated for each stimulus format separately (see Materials and Methods). Participants rated both stimulus formats with highly stable similarity judgments (split-half reliabilities: $r = 0.98$ for both dynamic and static stimuli). We used multidimensional scaling on the pairwise dissimilarities of each stimulus format to

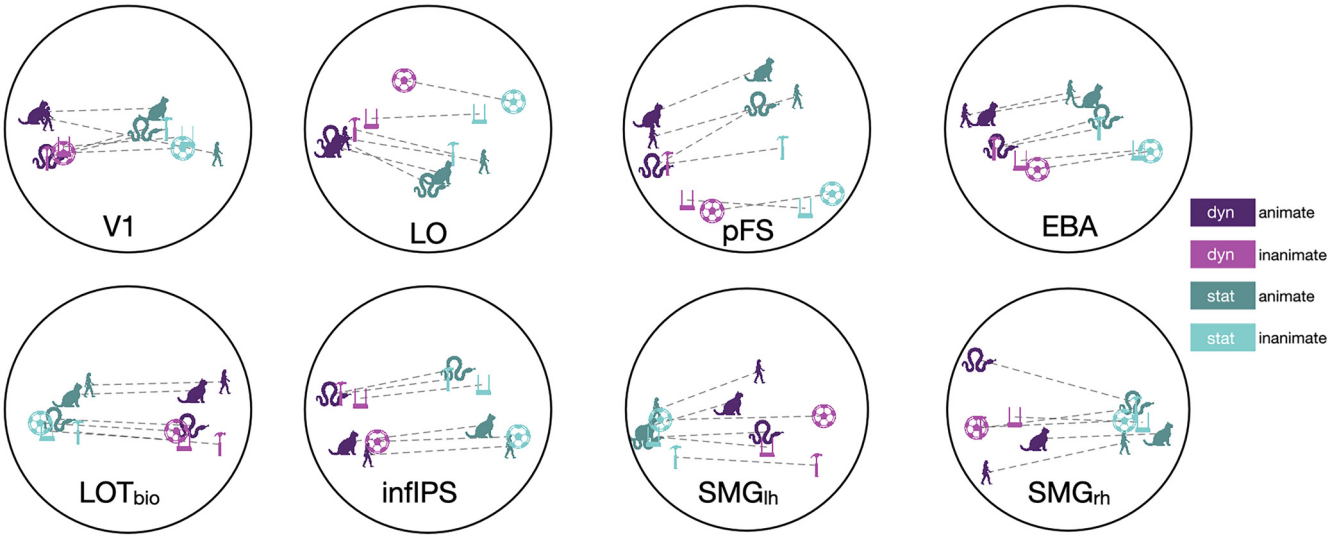

Figure 5. Multidimensional scaling visualization of fMRI response similarity between the object categories presented in the dynamic and static formats. MDS was performed on the similarity matrix obtained from the Euclidean distances of response patterns for the 12 conditions in each ROI. Dotted lines connect dynamic and static presentations of the same object category. Purple represents the dynamic condition. Green represents the static condition. Within each condition, the darker hues represent the animate categories, while the lighter hues represent the inanimate categories. The 6 object categories are symbolized with the following icons: human (person walking), mammal (cat), reptile (snake), tool (hammer), pendulum/swing (swing), and ball (soccer ball). Also, see Results from the LOT-atlas and STS-atlas regions defined by a probabilistic atlas in Extended Data Figure 2-1*E*, *F*.

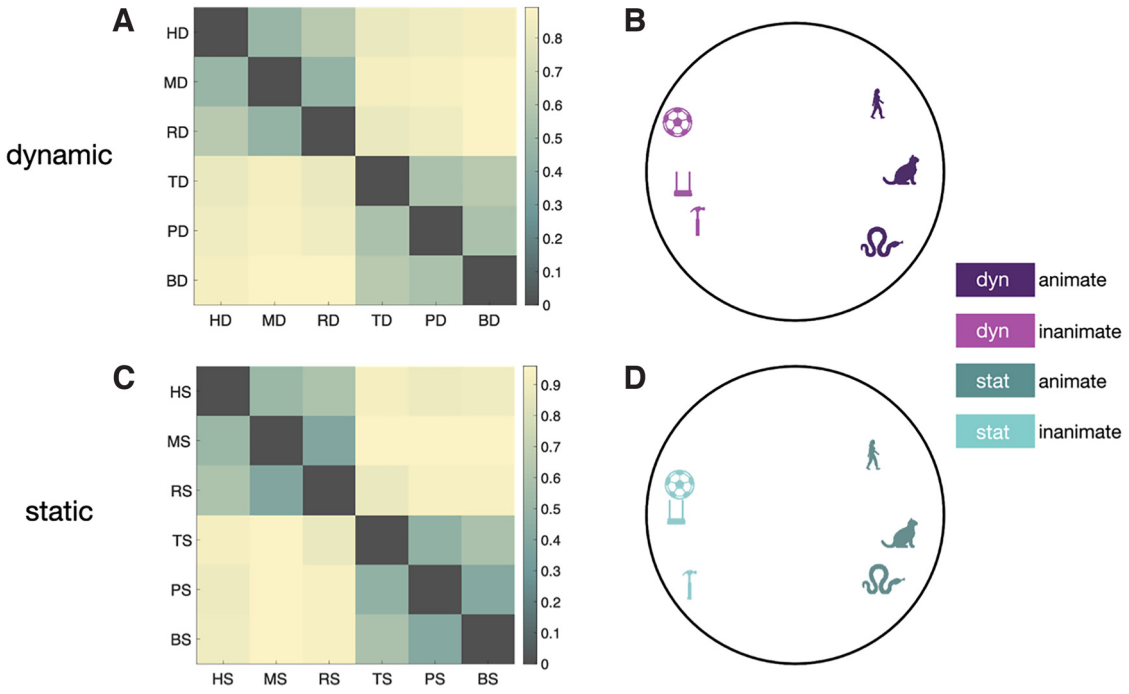

Figure 6. Odd-one-out similarity judgments of dynamic and static stimuli at the category level. Matrices represent pairwise dissimilarity scores between object categories in dynamic (***A***) and static (***C***) stimulus formats. Circle plots represent the object categories projected into the first two dimensions from multidimensional scaling on the dissimilarities in the dynamic (***B***) and static (***D***) stimuli. The 6 object categories are symbolized with the same icon designation as in Figure 5. Category abbreviations (dynamic/static): human (HD/HS), mammal (MD/MS), reptile (RD/RS), tool (TD/TS), pendulum/swings (PD/PS), and balls (BD/BS).

visualize the distance between object categories in the first two dimensions (Fig. 6*A*). The dynamic and static similarity judgments had highly similar structures, showing a clear separation between animate and inanimate categories in the first dimension. The animate (human, mammal, and reptile) and inanimate (tool, pendulum/swing, and ball) categories were also separated from each other along the second dimension for both types of stimuli. Overall, the dissimilarities from the dynamic and static tasks were highly correlated ($r = 0.98$, $p = 2.80\text{e-}10$).

To further explore the similarity structure of the dynamic and static stimuli at the exemplar level, a hierarchical clustering algorithm was used on the odd-one-out similarity judgments (Fig. 7). Similar to the MDS of odd-one-out judgments at the category level, a gross distinction between animate and inanimate objects was observed for both the static and dynamic conditions. Also,

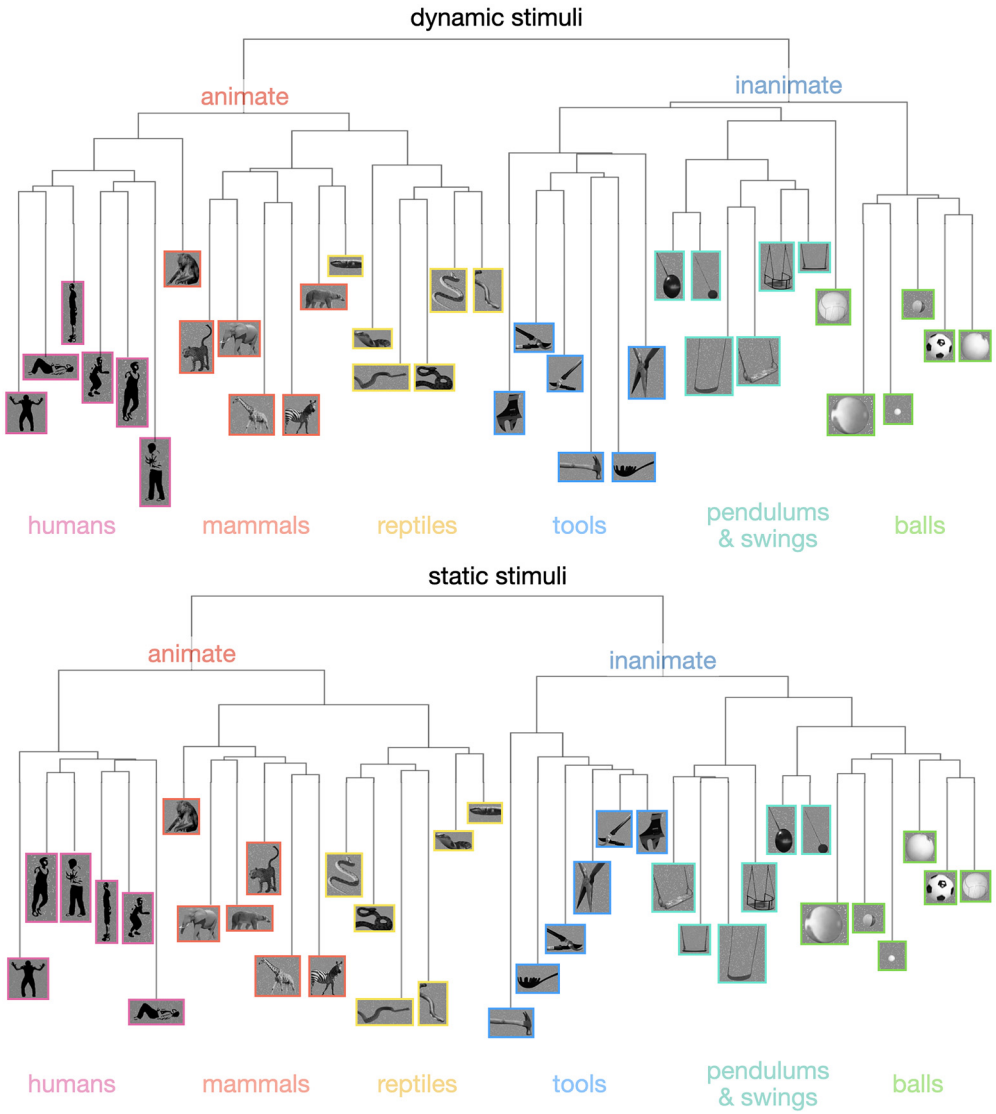

Figure 7. Hierarchical clustering of odd-one-out similarity judgments of the dynamic and static stimuli at the exemplar level. Edited versions of the static stimuli were used to visualize the similarity structure of both the dynamic (top) and static (bottom) stimuli, as the category of the dynamic stimuli cannot be gleaned from individual frames. The scale and position of the objects are not representative of the stimuli during presentation. Stimulus borders were colored to distinguish the 6 object categories: humans (pink), mammals (orange), reptiles (yellow), tools (blue), pendulums & swings (teal), and balls (green). The human stimulus examples were modified into two-tone images for this figure to de-identify the individuals in the stimuli.

the 3 object categories within the animate and inanimate superordinate categories are largely distinguished. We also see that the individual stimuli are primarily clustered within their object categories in both stimulus formats, suggesting that, when luminance-defined edges are not available, robust category information can be derived from dynamic motion-isolated inputs.

To investigate how the fMRI responses to object categories in each format relate to behavioral judgments of similarity, we correlated the dissimilarity scores from the dynamic and static behavioral experiments (split-half reliabilities: dynamic 0.985, static 0.985) to those obtained from the Euclidean distances between the multivariate response patterns in each ROI (LOSO reliabilities: dynamic 0.67 ± 0.12, static 0.61 ± 0.20). As shown in Figure 8, most ventral and lateral temporal regions (LO, pFS, EBA, LOT-biomotion) showed significant correlations with the object similarity judgments for both the dynamic and static stimuli (dynamic: r values > 0.25, p values < 0.007; static: r values > 0.15, p values < 0.05). These results replicate previous findings of robust correspondence between similarity judgments and fMRI responses to both static images (Cohen et al., 2017; Xu and Vaziri-Pashkam, 2019) and human body movements (Hafri et al., 2017; Yargholi et al., 2021) and demonstrate that the motion-defined representations of dynamic object category information show a similar correspondence in these regions. The responses in infIPS and right SMG were not correlated to object similarity judgments for either the dynamic or static stimuli (dynamic: r values < 0.13, p values > 0.11, static: r values < 0.02, p values > 0.58). While lack of correlation in these regions was unexpected, previous work has shown that dorsal regions have lower correlations with behavioral similarity judgments relative to ventral and lateral regions (Cohen et al., 2017; Xu and Vaziri-Pashkam, 2019). The activity in left SMG was significantly correlated with the similarity judgments for the dynamic stimuli ($r = 0.33$, $p = 0.001$), but not for the static stimuli ($r = 0.04$, $p = 0.59$). Similarly, the activity in V1 was significantly correlated with similarity judgments for the static stimuli ($r = 0.13$, $p = 0.02$), but not for the dynamic

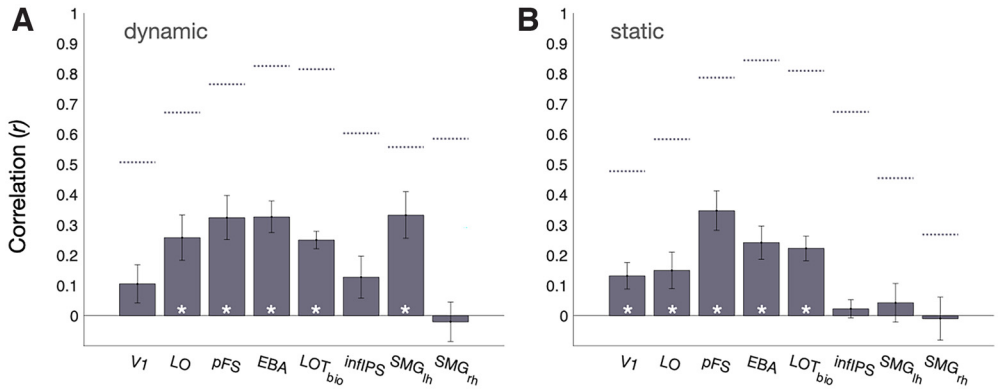

Figure 8. Correlation of Euclidean distance between multivariate fMRI responses and behavioral dissimilarity matrices for (***A***) dynamic and (***B***) static stimuli. $*p < 0.05$. Error bars indicate standard error of the mean. Dotted lines indicate the average noise ceiling for each ROI.

stimuli ($r = 0.11$, $p = 0.14$). The only significant difference between the correlations of the behavioral similarity judgments and the fMRI responses to the two conditions was found in the left SMG area, in which the correlation was significantly higher with similarity judgments of the dynamic stimuli compared with the static stimuli ($t_{(14)} = 3.32$, $p = 0.04$, Cohen's $d = 0.86$). These findings demonstrate that responses in the ventral and lateral region driven by both static and dynamic visual information contain robust object category information related to the object similarity judgments, while the left SMG only contains object similarity information when driven by its preferred stimulus format (dynamic). For regions such as infIPS and right SMG, the object category responses may be structured for a different purpose that is unrelated to subjective similarity judgments.

Discussion

Motion is an important visual cue that can provide category-relevant information in the absence of luminance-defined edges and form. Here, we introduce a novel approach to systematically separate form and motion signals and study the contribution of the motion signal to object category processing in isolation. To our knowledge, our study is the first to use this approach to compare the neural processing of form and motion signals from several animate and inanimate object categories. We sought to determine whether category-relevant information from the two sources is shared across the visual system by comparing dynamic and static category processing in ROIs across visual occipito-temporal and parietal cortices. The two highly dissimilar information sources produced distinct but sufficiently overlapping representations of animate and inanimate object categories to allow for across-format decoding in all higher-order ROIs.

Categorizing objects with motion information

We first tested whether the objects in our object kinematograms were recognizable with an online object identification task. Our behavioral study shows that not only do people accurately categorize motion-defined animate objects (Pinto, 1994, 2006; Pavlova et al., 2001), they also accurately categorize at least three motion-defined inanimate object categories. These results and our odd-one-out similarity task demonstrate that: (1) a wide range of animate and inanimate objects can be recognized from just motion information; and (2) people judge the similarity of objects from the two sources of information in a similar way.

Broadly speaking, two types of object information can be gleaned from motion cues: (1) structure from motion, that is, a percept of a form arising from the global integration of coherent local motion vectors; and (2) types of actions or features of actions that are diagnostic of a particular object, such as walking, fighting, tool use, bouncing, etc. Although systematically distinguishing these two sources was not within the scope of this study, both factors likely play an important role in subjects' judgments of object similarity. Novel object kinematograms that capture a greater range of actions from many viewpoints will be instrumental in teasing apart their relative contributions to object recognition from dynamic inputs.

Format-dependent processing of object categories

Our findings suggest that stimulus format matters for: (1) processing of animate and inanimate objects, indicated by the ROIs with significant interactions between stimulus format and univariate animacy preference (i.e., pFS and left SMG); and (2) discriminating object categories within-format, indicated by regions with significant differences in the multivariate classification accuracy of the responses to dynamic and static stimuli (i.e., all regions but infIPS). The most ventral and posterior regions (LO, EBA, and pFS) showed higher classification in the static condition, while the most dorsal and anterior regions (LOT-biomotion and bilateral SMG) had stronger classification in the dynamic condition. Interestingly, infIPS used both sources of information equivalently. Importantly, all ROIs but V1 showed robust responses to, and significant decoding accuracies of, all categories presented in both static image and dynamic motion formats. Thus, differential multivariate responses to object categories based on stimulus format in these regions are a matter of degree.

Animate and inanimate category processing from motion and form

In a study comparing responses to dynamic and static cues for humans and tools, Beauchamp et al. (2003) compared univariate fMRI responses between (1) full form videos and static images of humans and tools and (2) full form videos and PLDs of humans and tools. Beauchamp et al. (2003) argued for two processing pathways: form and motion. Lateral temporal regions (STS and MTG) respond to their preferred category, humans and tools, respectively, in both PLDs and videos, suggesting category preference from motion without requiring form. Meanwhile, ventral temporal cortex (lateral and medial fusiform) needed form information for category preference responses. Our results

demonstrate that this animacy preference topography is not exclusive to human and tool categories: it also expands to other animate objects, such as mammals and reptiles, and other inanimate objects, such as pendulums/swings and balls. It is possible that large-scale animacy preference maps (Konkle and Caramazza, 2013; Sha et al., 2015) found with static objects in the brain are also present for motion-defined stimuli.

Distinct but overlapping representations of object category for dynamic and static stimuli

We used a cross-classification approach to identify regions that have format independent responses. A similar analysis has been used previously to study fMRI responses to human actions in full form videos and images (Hafri et al., 2017). Our results are largely in qualitative agreement with those of Hafri et al. (2017), with the exception that we found significantly more widespread cross-classification, possibly because our static stimuli were source matched to our dynamic stimuli. Cross-decoding in all regions, apart from V1, suggests that the object category representations driven by static and dynamic information were sufficiently distinct to allow for significant within-format classification, but also sufficiently overlapping that their shared information could lead to significant cross-classification. These results suggest the existence of abstract object category responses that pool information about object category across various cues in the visual input, in line with previous reports of cross-modal integration in visual cortex in neurotypical and congenitally blind individuals (Fairhall and Caramazza, 2013; Peelen et al., 2014; Kumar et al., 2017).

Relationship between brain and behavior

Multivariate responses to the dynamic and static conditions in the ventral and lateral regions (LO, pFS, EBA, and LOT-biomotion) were correlated with the object similarity judgments of the dynamic and static stimuli, respectively. There were no differences in correlation between the two formats. This implies that the fMRI responses in these regions follow the structure of the stimulus similarity characterized by our odd-one-out experiment. The only region to show a difference in correlation across the stimulus conditions was the left SMG, which showed higher correlations for the fMRI responses to the dynamic stimuli. By contrast, the fMRI responses in the right SMG showed no significant correlation with behavioral judgments for either condition, indicating lateralization of category processing within the supramarginal area to the left hemisphere. This left lateralization has been shown previously in research on tool processing (Beauchamp et al., 2003; Peeters et al., 2013; Rüther et al., 2014) and in studies of modality-independent category processing across images and words (Fairhall and Caramazza, 2013; Wurm and Caramazza, 2019). Importantly, not all regions that showed significant animacy preference or object category decoding had responses that were significantly correlated with the similarity structure of the behavioral judgments. In V1 and infIPS, the fMRI responses to both conditions were unrelated to the similarity judgments of both stimulus types, suggesting that these regions were extracting features irrelevant to similarity judgments on the objects.

Contributions of the dorsal, ventral, and lateral visual pathways

Giese and Poggio (2003) have suggested segregation between processing of motion and form in the visual system. They argue that motion is predominantly processed in traditionally "dorsal" regions, while form is reserved for the ventral stream (Giese and Poggio, 2003). In a recent review, Pitcher and Ungerleider (2021) have suggested the presence of a third "lateral" motion-sensitive visual pathway for social visual processing (i.e., for the perception of face and body movements). In line with these theories, we found more information for objects in the static condition in the ventral pFS region and in the dynamic condition in lateral LOT-biomotion and dorsal SMG regions. However, above chance decoding for both static and dynamic conditions in all regions argues against strict segregation of the processing of motion from form. One possible reason we observe significant category decoding from motion cues in the ventral stream is that both dorsal/lateral and ventral pathways receive similar inputs from early motion sensitive areas, such as V3 (McLeod et al., 1996; Ho and Giaschi, 2009). Another explanation is that bidirectional connectivity between dorsal/lateral and ventral pathways supports the transfer of task relevant information (O'Toole et al., 2002; Bernstein and Yovel, 2015; Freud et al., 2018; Collins et al., 2019; Ayzenberg and Behrmann, 2022). While we cannot distinguish between these possibilities, our within-dynamic and across-format decoding results demonstrate that motion-derived object category information is not limited to dorsal/lateral regions.

Despite our emphasis on the distributed processing of dynamic object category cues in the two or three pathways, the neuropsychological literature suggests that each contributes something different to object category processing from dynamic cues. For example, patients with damage to the ventral form pathway might show deficits in structure from motion perception but intact PLD recognition (Gilaie-Dotan et al., 2013, 2015). These studies suggest that, while the form pathway is not strictly necessary for recognition of PLDs, it could facilitate extracting form from complex motion to create more robust multiview representations of objects. Furthermore, case studies have demonstrated that lower-level motion performance (e.g., motion identification and coherence) can be impaired without corresponding complete impairment to the recognition of PLDs, a higher-level motion task (Vaina et al., 1990; Mcleod et al., 1996). In addition, it has been shown that damage to dorsal parietal cortex could lead to deficits in visual search for biological motion without an impairment in form from motion (Battelli et al., 2003). These results suggest distinct roles of regions in ventral and dorsal/lateral pathways in the recognition of our dynamic stimuli, but additional experiments are required to confirm these conjectures.

In conclusion, our study demonstrates that, in regions across occipito-temporal and parietal cortices, category responses driven by isolated motion signals parallel category responses to static form signals in a number of ways. Regions that are traditionally considered part of the visual object recognition pathway that processes static information also extract robust object category information from isolated motion signals. Indeed, object category information from static and dynamic signals overlaps. Future studies can further probe the nature of motion-defined object representations by generating kinematograms of a larger set of objects performing different movements from multiple viewpoints. Such studies will be important for furthering our understanding of how, in a dynamic scene with multiple objects, various visual cues to object category are processed and integrated together to form rich and robust object representations in the human brain.

References

Ayzenberg V, Behrmann M (2022) The dorsal visual pathway represents object-centered spatial relations for object recognition. J Neurosci 42:4693–4710.

Barclay CD, Cutting JE, Kozlowski LT (1978) Temporal and spatial factors in gait perception that influence gender recognition. Percept Psychophys 23:145–152.

Bassili JN (1978) Facial motion in the perception of faces and of emotional expression. J Exp Psychol Hum Percept Perform 4:373–379.
Battelli L, Cavanagh P, Thornton IM (2003) Perception of biological motion in parietal patients. Neuropsychologia 41:1808–1816.
Beauchamp MS, Lee KE, Haxby JV, Martin A (2003) fMRI responses to video and point-light displays of moving humans and manipulable objects. J Cogn Neurosci 15:991–1001.
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B (Methodological) 57:289–300.
Bernstein M, Yovel G (2015) Two neural pathways of face processing: a critical evaluation of current models. Neurosci Biobehav Rev 55:536–546.
Bonda E, Petrides M, Ostry D, Evans A (1996) Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J Neurosci 16:3737–3744.
Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436.
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.
Cavanagh P, Mather G (1989) Motion: the long and short of it. Spat Vis 4:103–129.
Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions Intel Syst Technol 2:1–27.
Cohen MA, Alvarez GA, Nakayama K, Konkle T (2017) Visual search for object categories is predicted by the representational architecture of high-level visual cortex. J Neurophysiol 117:388–402.
Collins E, Freud E, Kainerstorfer JM, Cao J, Behrmann M (2019) Temporal dynamics of shape processing differentiate contributions of dorsal and ventral visual pathways. J Cogn Neurosci 31:821–836.
Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.
Cutting JE, Kozlowski L (1977) Recognition of friends by their walk. Bull Psychon Soc 9:353–356.
Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980.
Dittrich WH, Troscianko T, Lea SE, Morgan D (1996) Perception of emotion from dynamic point-light displays represented in dance. Perception 25:727–738.
Engell AD, McCarthy G (2013) Probabilistic atlases for face and biological motion perception: an analysis of their reliability and overlap. Neuroimage 74:140–151.
Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558.
Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image Analysis, pp 363–370. Berlin: Springer.
Freud E, Robinson AK, Behrmann M (2018) More than action: the dorsal pathway contributes to the perception of 3-D structure. J Cogn Neurosci 30:1047–1058.
Furl N, Hadj-Bouziane F, Liu N, Averbeck BB, Ungerleider LG (2012) Dynamic and static facial expressions decoded from motion-sensitive areas in the macaque monkey. J Neurosci 32:15952–15962.
Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience 4:179–192.
Giese MA (2013) Biological and body motion perception. The Oxford handbook of perceptual organization, pp 575–596. Oxford, England:Oxford University Press.
Gilaie-Dotan S, Saygin AP, Lorenzi LJ, Egan R, Rees G, Behrmann M (2013) The role of human ventral visual cortex in motion perception. Brain 136:2784–2798.
Gilaie-Dotan S, Saygin AP, Lorenzi LJ, Rees G, Behrmann M (2015) Ventral aspect of the visual form pathway is not critical for the perception of biological motion. Proc Natl Acad Sci USA 112:E361–E370.
Grossman ED, Blake R (2002) Brain areas active during visual perception of biological motion. Neuron 35:1167–1175.
Hafri A, Trueswell J, Epstein R (2017) Neural representations of observed actions generalize across static and dynamic visual input. J Neurosci 37:3056–3071.
Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57:243–259.
Hirai M, Hiraki K (2006) The relative importance of spatial versus temporal structure in the perception of biological motion: an event-related potential study. Cognition 99:B15–B29.
Ho CS, Giaschi DE (2009) Low- and high-level first-order random-dot kinematograms: evidence from fMRI. Vision Res 49:1814–1824.
Johansson G (1973) Visual perception of biological motion and a model of its analysis. Percept Psychophys 14:201–211.
Johansson G (1976) Spatio-temporal differentiation and integration in visual motion perception. Psychol Res 38:379–393.
Kaiser MD, Shiffrar M, Pelphrey KA (2012) Socially tuned: brain responses differentiating human and animal motion. Soc Neurosci 7:301–310.
Kleiner M, Brainard D, Pelli D (2007) What's new in Psychtoolbox-3? Perception 36 (ECVP Abstr Suppl):14.
Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. J Neurosci 33:10235–10242.
Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis-connecting the branches of systems neuroscience. Front Syst Neurosci 4:1–28.
Kumar MK, Federmeier KD, Fei-Fei L, Beck DM (2017) Evidence for similar patterns of neural activity elicited by picture- and word-based representations of natural scenes. Neuroimage 155:422–436.
Kundu P, Inati SJ, Evans JW, Luh WM, Bandettini PA (2012) Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. Neuroimage 60:1759–1770.
Mather G, West S (1993) Recognition of animal locomotion from dynamic point-light displays. Perception 22:759–766.
McLeod P, Dittrich W, Driver J, Perrett D, Zihl J (1996) Preserved and impaired detection of structure from motion by a 'motion-blind' patient. Visual Cognition 3:363–392.
Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang XR, Just M, Newman S (2004) Learning to decode cognitive states from brain images. Machine Learning 57:145–175.
Mitkin AA, Pavlova MA (1990) Changing a natural orientation: recognition of biological motion pattern by children and adults. Psychologische Beitrage 32:28–35.
Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.
Nunnally JC (1970) Introduction to psychological measurement. New York: McGraw-Hill.
O'Toole AJ, Roark DA, Abdi H (2002) Recognizing moving faces: a psychological and neural synthesis. Trends Cogn Sci 6:261–266.
Op de Beeck HP, Torfs K, Wagemans J (2008) Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. J Neurosci 28:10111–10123.
Papeo L, Wurm MF, Oosterhof NN, Caramazza A (2017) The neural representation of human versus nonhuman bipeds and quadrupeds. Sci Rep 7:1–8.
Pavlova M, Krägeloh-Mann I, Sokolov A, Birbaumer N (2001) Recognition of point-light biological motion displays by young children. Perception 30:925–933.
Pavlova M, Lutzenberger W, Sokolov A, Birbaumer N (2004) Dissociable cortical processing of recognizable and non-recognizable biological movement: analysing gamma MEG activity. Cereb Cortex 14:181–188.
Peelen MV, He C, Han Z, Caramazza A, Bi Y (2014) Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. J Neurosci 34:163–170.
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.
Peeters R, Rizzolatti G, Orban GA (2013) Functional properties of the left parietal tool use region. Neuroimage 78:83–93.
Peissig JJ, Tarr MJ (2007) Visual object recognition: do we know more now than we did 20 years ago? Annu Rev Psychol 58:75–96.
Pereira F, Botvinick M (2011) Information mapping with pattern classifiers: a comparative study. Neuroimage 56:476–496.
Pinto J (1994) Human infants' sensitivity to biological motion in pointlight cats. Infant Behav Dev 17:871.
Pinto J (2006) Developing body representations: a review of infants' responses to biological-motion displays. In: Perception of the human

body from the inside out (Knoblich G, Grosjean M, Thornton I, Shiffrar M, eds), pp 305–322. Oxford, England:Oxford University Press.
Pinto J, Shiffrar M (2009) The visual perception of human and animal motion in point-light displays. Soc Neurosci 4:332–346.
Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, Kanwisher N (2011) Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56:2356–2363.
Pitcher D, Ungerleider LG (2021) Evidence for a third visual pathway specialized for social perception. Trends Cogn Sci 25:100–110.
Posse S, Wiese S, Gembris D, Mathiak K, Kessler C, Grosse-Ruyken ML, Elghahwagi B, Richards T, Dager SR, Kiselev VG (1999) Enhancement of BOLD-contrast sensitivity by single-shot multi-echo functional MR imaging. Magn Reson Med 42:87–97.
Ptito M, Faubert J, Gjedde A, Kupers R (2003) Separate neural pathways for contour and biological-motion cues in motion-defined animal shapes. Neuroimage 19:246–252.
Rizzolatti G, Craighero L (2004) The mirror-neural system. Annu Rev Neurosci 27:169–192.
Rüther NN, Tettamanti M, Cappa SF, Bellebaum C (2014) Observed manipulation enhances left fronto-parietal activations in the processing of unfamiliar tools. PLoS One 9:e99401.
Saygin AP, Wilson SM, Hagler DJ, Bates E, Sereno MI (2004) Point-light biological motion perception activates human premotor cortex. J Neurosci 24:6181–6188.
Schenk T, Zihl J (1997) Visual motion perception after brain damage: II. Deficits in form-from-motion perception. Neuropsychologia 35:1299–1310.
Scholl BJ, Gao T (2013) Perceiving animacy and intentionality: visual processing or higher-level judgment. Soc Percept 4629:197–229.
Schultz J, Bülthoff HH (2013) Parametric animacy percept evoked by a single moving dot mimicking natural stimuli. J Vis 13:15.
Sha L, Haxby JV, Abdi H, Guntupalli JS, Oosterhof NN, Halchenko YO, Connolly AC (2015) The animacy continuum in the human ventral vision pathway. J Cogn Neurosci 27:665–678.
Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. Science 210:390– 398.
Tootell RB, Reppas JB, Dale AM, Look RB, Sereno MI, Malach R, Brady TJ, Rosen BR (1995) Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. Nature 375:139–141.
Vaina LM, Lemay M, Bienfang DC, Choi AY, Nakayama K (1990) Intact 'biological motion' and 'structure from motion' perception in a patient with impaired motion mechanisms: a case study. Vis Neurosci 5:353–369.
Vaziri-Pashkam M, Xu Y (2017) Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. J Neurosci 37:8767–8782.
Vaziri-Pashkam M, Xu Y (2019) An information-driven 2-pathway characterization of occipitotemporal and posterior parietal visual object representations. Cereb Cortex 29:2034–2050.
Vaziri-Pashkam M, Taylor J, Xu Y (2019) Spatial frequency tolerant visual object representations in the human ventral and dorsal visual processing pathways. J Cogn Neurosci 31:49–63.
Virtanen P, et al., SciPy 1.0 Contributors (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272.
Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4:274–290.
Watson JD, Myers R, Frackowiak RS, Hajnal JV, Woods RP, Mazziotta JC, Shipp S, Zeki S (1993) Area V5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. Cereb Cortex 3:79–94.
Wurm MF, Caramazza A (2019) Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. Nat Commun 10:289.
Xu Y, Chun MM (2009) Selecting and perceiving multiple visual objects. Trends Cogn Sci 13:167–174.
Xu Y, Vaziri-Pashkam M (2019) Task modulation of the 2-pathway characterization of occipitotemporal and posterior parietal visual object representations. Neuropsychologia 132:107140.
Yargholi E, Hossein-Zadeh GA, Vaziri-Pashkam M (2021) Two distinct networks containing position-tolerant representations of actions in the human brain. bioRxiv. https://doi.org/10.1101/2021.06.17.448825.
Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8:665–670.
Zeki S, Watson JD, Lueck CJ, Friston KJ, Kennard C, Frackowiak RS (1991) A direct demonstration of functional specialization in human visual cortex. J Neurosci 11:641–649.