Behavioral/Cognitive

General Transformations of Object Representations in Human Visual Cortex

Emily J. Ward,[1] Leyla Isik,[2] and Marvin M. Chun[3]
[1]Department of Psychology, University of Wisconsin–Madison, Madison, Wisconsin 53706, [2]Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and [3]Yale University, New Haven, Connecticut 06520

The brain actively represents incoming information, but these representations are only useful to the extent that they flexibly reflect changes in the environment. How does the brain transform representations across changes, such as in size or viewing angle? We conducted a fMRI experiment and a magnetoencephalography experiment in humans (both sexes) in which participants viewed objects before and after affine viewpoint changes (rotation, translation, enlargement). We used a novel approach, representational transformation analysis, to derive transformation functions that linked the distributed patterns of brain activity evoked by an object before and after an affine change. Crucially, transformations derived from one object could predict a postchange representation for novel objects. These results provide evidence of general operations in the brain that are distinct from neural representations evoked by particular objects and scenes.

Key words: object recognition; perception; transformation; viewpoint invariance; visual representations

Significance Statement

The dominant focus in cognitive neuroscience has been on how the brain represents information, but these representations are only useful to the extent that they flexibly reflect changes in the environment. How does the brain transform representations, such as linking two states of an object, for example, before and after an object undergoes a physical change? We used a novel method to derive transformations between the brain activity evoked by an object before and after an affine viewpoint change. We show that transformations derived from one object undergoing a change generalized to a novel object undergoing the same change. This result shows that there are general perceptual operations that transform object representations from one state to another.

Introduction

The primary function of the brain is to adaptively interact with the world, which requires both representing and transforming incoming information (deCharms and Zador, 2000). Representations are information-bearing structures, and transformations are the computational procedures that operate on those structures. The dominant focus in cognitive neuroscience has been on neural representation, such as how the visual system codes for features, objects, and scenes. Patterns of brain activity (measured with fMRI and other imaging modalities) reflect how neuronal populations code and represent information (Kriegeskorte et al., 2008b). But these representations are only useful to the extent that they flexibly reflect changes in the environment and behavioral goals. For example, we need to recognize an umbrella regardless of whether it is closed or open, but we also need to know that it is only useful to provide protection against rain when it is open. How does the brain link object state representations, such as before and after an object undergoes a change?

Some of the most common changes that our brain must link are affine visual changes, such as when we move toward or away from an object. When we see an object from a new viewpoint, there are two problems our perceptual system needs to solve: (1) recognizing that the object is the same, despite the change; and (2) recognizing how the viewpoint has changed, regardless of the object. The first problem of viewpoint invariance has been studied widely (DiCarlo and Cox, 2007) and is demonstrated by similarity between object representations (as assessed by distance metrics) across viewpoints. But it is unknown whether the second problem is solved in the same similarity-based manner or whether there are general transformations of object representations. For example, a transformation derived from one object

Received Sept. 27, 2017; revised Aug. 7, 2018; accepted Aug. 9, 2018.

Author contributions: E.J.W. wrote the first draft of the paper; E.J.W., L.I., and M.M.C. edited the paper; E.J.W., L.I., and M.M.C. designed research; E.J.W. and L.I. performed research; E.J.W. analyzed data; E.J.W., L.I., and M.M.C. wrote the paper.

We thank Adam Chekroud, Matthew Johnson, Gary Lupyan, Brian Scholl, and the members of the Yale Visual Cognitive Neuroscience Laboratory for helpful conversation and/or comments on earlier drafts.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Emily J. Ward, Department of Psychology, University of Wisconsin–Madison, 1202 W. Johnson Street, Madison, WI 53706. E-mail: ejward@wisc.edu.

DOI:10.1523/JNEUROSCI.2800-17.2018

undergoing a change should generalize to a novel object undergoing the same change: given an object's initial prechange representation, it should be possible to predict its final postchange representation. Furthermore, the transformations should operate on object representations that are high-level and generalizable to novel stimuli, and not on "image-like" patterns (e.g., the stimuli themselves or retinotopic representations).

Alternatively, the brain may link perceptual states simply through similarity. Patterns of brain activity can be systematically characterized through distance metrics (e.g., representational similarity analysis) (Kriegeskorte et al., 2008a), and predictable state changes between stimuli result in higher pattern similarity in the medial temporal lobe among the stimuli (Hindy and Turk-Browne, 2016). However, if the brain also links perceptual states through transformations beyond their initial similarity, this could better account for differences between the representations (e.g., predicting a pattern that is a better match to the true postchange representation) and, importantly, would generalize to new objects.

To determine how object representations are linked from one state to another, we looked for transformations that operate whenever we see any object undergoing a common viewpoint change. In the first experiment, we scanned participants with fMRI while they viewed images of three different objects before and after one of five affine changes. We then used a novel approach, representational transformation analysis, to derive a transformation function that could be applied to the prechange object representation to generate the postchange object representation (like a function that transforms x [prechange] into y [postchange]: $f(x) = y$). This idea is similar to the classic idea of visual routines (Ullman, 1984), which are primitive visual processing operations that can establish properties and relationships that are not explicit in the initial representations. In the second experiment, we analyzed MEG data (Isik et al., 2014) from participants who had viewed six different objects at four different viewpoints. Using the same representational transformation analysis, we looked for evidence of distinct computational stages encompassed by the transformation. Across these two studies, we find general transformations between representations of objects that have undergone a common viewpoint change, such that when the transformation is applied to a novel representation, it produces an accurate prediction of a new representation.

Materials and Methods

Experiment 1: representational transformation analysis with fMRI

Participants. Fourteen participants (6 men and 8 women) were recruited from the Yale University community. The Yale University Human Subjects Committee approved the experimental protocol. Participants provided informed written consent before the experiment. Participants had normal or corrected-to-normal vision. Participants were paid for their participation. The number of participants was selected to be sufficient for multivoxel pattern detection using brief fMRI acquisition runs (Coutanche and Thompson-Schill, 2012).

Apparatus. fMRI data were acquired at the Magnetic Resonance Research Center at Yale University on a 3-T Trio (Siemens) equipped with a 32-channel head coil. T1-weighted anatomical images were acquired using a 3D MPRAGE sequence (TR = 2530 ms, TE = 2.77 ms, time to inversion = 1100 ms, voxel size = 1 × 1 × 1 mm, matrix size = 256 × 256 × 256). T2*-weighted images sensitive to BOLD contrasts were acquired using a multiband EPI sequence (TR = 1000 ms, TE = 30 ms, flip angle = 62°, acquisition matrix = 84 × 84, in-plane resolution = 2.5 mm square, 51 axial-oblique slices parallel to the anterior and posterior commissural line, slice thickness = 2.5, multiband 3, acceleration factor = 2). Stimuli were presented using PsychoPy (Peirce, 2007) and displayed through an LCD projector on a rear-projection screen. Responses were recorded using a 2-button fiberoptic response pad system.

Stimuli. Stimuli were color photographs of a single exemplar from three distinct object categories: a moth (natural), a rain boot (manmade), and a greeble-like object (novel) (Gauthier et al., 2003) presented on a white background. The initial images were ~10° × 10° of visual angle in size and were oriented in a canonical viewpoint. The initial images were considered the "prechange" state of the object. For the "postchange" state, the objects were subjected to five different affine changes: identity (no change), size increase (by 50%), size decrease (by 50%), left in-plane rotation (60° to the left), and right in-plane rotation (−60° to the right). There were thus 15 images in total (3 stimulus types and 5 change types). All images were presented in isolation at the center of the display.

Experimental design and procedure. Images were presented for 300 ms with 3966 ms interstimulus interval with no jitter. The prechange and postchange images were always presented sequentially, but the stimulus type (natural, manmade, and novel) and change type (identical, size increase, size decrease, left rotation, right rotation) were randomized across runs and across participants.

There were 15 runs, each lasting 3 min and 12 s. Brief runs of this type have been demonstrated to improve pattern classification (Coutanche and Thompson-Schill, 2012). On each run, each state for each stimulus type and each change type was presented once. In addition, there were 5 trials on which an image (drawn randomly from 15 images types) "jiggled." Participants were instructed to look for these jiggles and press a button whenever they saw it. To aid in subsequent fMRI modeling, 10 trials were null trials on which nothing appeared on the screen except a fixation cross. In total, there were 45 trials (3 stimuli × 5 changes × 2 states [prechange and postchange] + 5 jiggle trials + 10 null trials = 45 trials).

fMRI data analysis. Functional images were corrected for motion and for differences in slice timing. Data were registered to MNI152 standard space template (2 mm isotropic voxel size). Voxels were smoothed using a 5 mm FWHM Gaussian filter (e.g., Xue et al., 2010), which may reduce noise and improve sensitivity after motion correction (Kamitani and Sawahata, 2010). Cortical reconstruction of each participant's data was performed with FreeSurfer software (http://surfer.nmr.mgh.harvard.edu) (Fischl, 2012). For each run, a quadratic polynomial was fit and removed to eliminate drift. Data were analyzed using FSL (Smith et al., 2004) and in-house Python scripts.

Our analyses were run on lateral occipital ROIs defined in standard space on a subject-by-subject basis from their FreeSurfer cortical reconstruction. All transformation analyses were run on the results of an initial GLM. The GLM was fit separately for each of the 15 runs. In each GLM, each trial served as a separate regressor (3 stimuli × 5 changes × 2 states [prechange and postchange] + 5 jiggle trials = 35 total regressors in the GLM). Although the jiggle trials were included in the GLM, they were not included in further analysis. Hemodynamic response function was fit to the 300 ms stimulus presentation using a double gamma function with $\sigma = 2.449$ and delay of 6 s for the first gamma, and $\sigma = 4$ and delay of 16 s for the second gamma. This GLM resulted in β values that were extracted for each trial and for each voxel within the lateral occipital anatomical ROI.

Representational transformation analysis. Transformation matrices were computed separately for each participant, for each object, and for each affine change. fMRI data were segmented into 15 folds, corresponding to 15 experimental scan runs: 14 training folds and 1 left-out testing fold. This approach is a type of cross-validation, where the transformation matrix can be generated on a subset of the data and then evaluated on the held-out test data.

One stimulus was selected as the training stimulus, and one stimulus was selected as the validation stimulus. In the training folds, the data for the training stimulus were separated into "prechange" and "postchange" patterns based on whether they reflected the pattern evoked by the pattern evoked by the mid-sized, unrotated object or the pattern evoked by the object that had undergone an affine change.

Each voxel (or channel, for Experiment 2) composing the prechange patterns was used as an individual predictor, [X_1, X_2, X_3, $X_4 \ldots X_n$], and

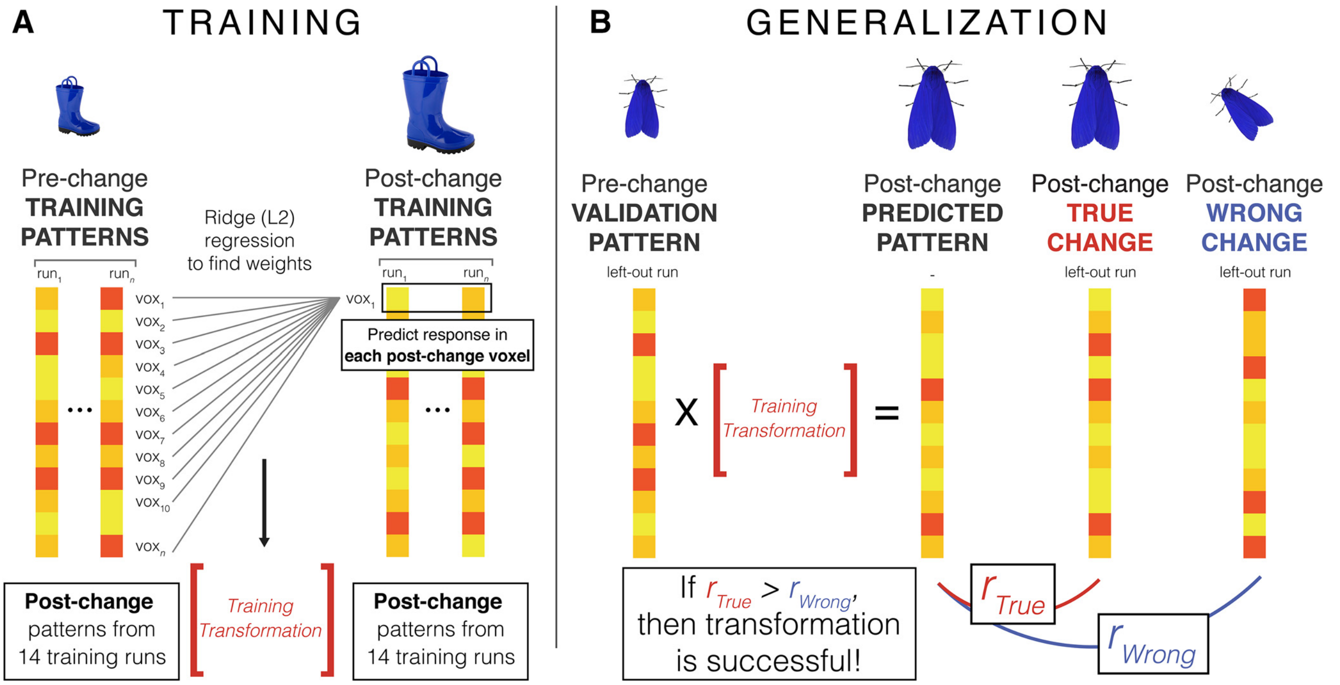

Figure 1. Overview of the representational transformation analysis. ***A***, Transformation matrices were obtained by finding the relationship between prechange patterns and postchange patterns for a training stimulus using L2-regularized regression. ***B***, The transformation matrix was then generalized to a new object: the transformation was applied to a held-out prechange pattern for a new object, generating a postchange predicted pattern. This pattern was correlated with the True Change pattern and the Wrong Change pattern. A higher correlation between the predicted pattern and the true pattern indicated that the transformation was successful.

each voxel composing the postchange patterns was used as an individual dependent variable [Y_1, Y_2, Y_3, Y_4 . . . Y_n] (Fig. 1*A*). The goal was to predict the response in each voxel in the postchange state (e.g., Y_1) from the response in all voxels in the prechange state (e.g., [X_1, X_2, X_3, X_4 . . . X_n]). In principle, this goal can be accomplished by iteratively fitting linear regressions to predict each voxel in the postchange state from all the voxels in the prechange state. In practice, a more efficient and equivalent approach is to compute all models simultaneously by including all data in one linear regression, which is possible because the same predictors are used in all models.

A linear model was fit with an n [voxel] × 14 [runs] matrix of predictors, X (and manually fit intercept) to predict an n [voxels] × 14 [runs] matrix of dependent variables, Y. The solution to this linear regression is as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where $\hat{\beta}$ is an $n \times n$ matrix of values reflecting how much each voxel in the prechange state should be weighted to predict each voxel in the postchange state. Because the number of predictors (voxels) was much more than the number of samples (runs), it was necessary to regularize the model. We used an L2 regularization to our linear regression (i.e., ridge regression), which penalizes the sum of squares of the weights. The solution to this ridge regression is as follows:

$$\hat{\beta} = (X^T X + \alpha I)^{-1} X^T Y$$

where α is the ridge regression penalty (set as $\alpha = 1$ in both experiments) and I is the identity matrix. $\hat{\beta}$ served as the training transformation matrix.

Once this transformation matrix was computed, it was applied to the prechange pattern of the test stimulus in the remaining 15th testing fold by taking the dot product of the n [voxels] × 1 [run] matrix of predictors and the $n \times n$ transformation matrix to yield an n [voxels] × 1 [run] predicted postchange pattern (Fig. 1*B*).

We assessed the success of the transformation by correlating the predicted postchange pattern to the true postchange pattern (e.g., did our predicted pattern for "big moth" match the actual pattern for "big moth?"). Thus, the higher the correlation between the predicted postchange pattern and the true postchange pattern, the better the transformation matrix was able to capture the general relationship between the prechange and postchange patterns.

This process was repeated for the following: (1) each object, ensuring that all objects served as the test stimulus; and (2) each affine change, ensuring that all changes types served as the training transformation (Fig. 1), for a total of 15 (3 objects × 5 affine changes) 15-fold training procedures.

Control analyses and permutation testing. To evaluate the correlations, we conducted four control analyses that produced comparison correlation values: (1) Wrong Change—for each predicted postchange pattern, we correlated it with each wrong postchange pattern for the same object; (2) Wrong Object—for each predicted postchange pattern, we correlated it with the same postchange pattern for each wrong object; (3) Mismatched Labels—for each training fold, we permuted the labels on the prechange and postchange patterns before finding the transformation matrix; and (4) Scrambled Transformation—for each transformation matrix, we permuted the coefficients before applying it to the test prechange pattern.

The primary control analysis was the Wrong Change analysis, which tested whether the transformation was truly specific to the affine change of interest or just producing object information (e.g., does the predicted pattern for "big moth" resemble the pattern for "any moth" because the representations are completely invariant?). We correlated the predicted postchange pattern to the postchange patterns from each of the wrong affine changes (Wrong Change) (e.g., we correlated the predicted pattern for big boot to the pattern for left rotated boot, etc.). This process was repeated for each affine change, ensuring that patterns evoked by all types of affine changes served as the true and wrong patterns. In order for the transformation to have succeeded, the predicted postchange pattern should better match the true pattern (higher correlation) than the wrong patterns (lower correlation).

The secondary control analyses were also informative: the Wrong Object control tested whether the transformation was specific to the object of interest or just generating patterns based on low-level features (e.g., does the predicted pattern for "big moth" resemble the pattern for "big anything" because all big items are represented with more retinotopic area?). We correlated the predicted postchange pattern to the postchange patterns for correct affine change but the wrong object (Wrong Object) (e.g., we correlated the predicted pattern for big boot to the pattern for big moth). This process was repeated for each stimulus, ensuring that patterns evoked by all objects served as the true and wrong patterns. In order for the transformation to have succeeded, the predicted postchange pattern should better match the true pattern (higher correlation) than the wrong patterns (lower correlation).

We also conducted two permutation analyses to approximate chance performance: a label permutation approach (Mismatched Labels) and a coefficient permutation approach (Scrambled Transformation). The Mismatched Label and Scrambled Transformation analyses ensured the quality of the regression analyses and integrity of the patterns. In the Mismatched Label permutation approach, we used the same procedure as described above with the following exceptions: In the training folds for a given stimulus, the affine-change labels and state (prechange or postchange) labels were randomly shuffled ($n = 1000$) before the data were separated into "prechange" and "postchange" patterns. The transformation matrix was generated based on these permuted labels and then evaluated on the held-out test data (which retained the original, correct affine-change and state labels). Given that there was a random relationship between the prechange and postchange training patterns, the resulting transformation matrix should produce a mostly random postchange predicted pattern, although some stimulus identity information may be preserved because the labels were permuted within stimulus.

In the Scrambled Transformation permutation approach, we used the representation transformation analysis procedure with the following exception: after the transformation matrix was computed, the matrix of coefficients was randomly shuffled (between and among rows) ($n = 1000$) before it was applied to a validation prechange pattern. Given that transformation matrix would now give a random mapping between the prechange and postchange voxels, it should produce a random postchange predicted pattern. Thus, the correlation between the true postchange pattern and the random postchange predicted pattern should also be at chance.

Transformation similarity versus initial pattern similarity. Another important factor to consider was whether the success of our transformations depended heavily on initial pattern similarity between the prechange and true postchange patterns. If the transformations were simply approximating initial pattern similarity (e.g., due to temporal correlation or representational similarity), then the transformations would be analogous to an identity function (like multiplying by 1), and representational transformation analysis would serve no purpose beyond pattern similarity analyses. Therefore, it was critical that the transformation analysis produce a pattern that is a better match (i.e., more correlated) to the true postchange pattern than the initial similarity between the prechange and postchange patterns.

We computed the Pearson correlation between prechange versus postchange patterns (i.e., initial pattern similarity, $r_{pattern\ similarity}$) and compared it with the Pearson correlation between the predicted postchange pattern and the true postchange pattern (i.e., the True Change transformation prediction similarity, $r_{transformation\ similarity}$). If $r_{transformation\ similarity}$ is greater than $r_{pattern\ similarity}$, the transformation produced a predicted postchange pattern that is more similar to the true postchange pattern than the prechange and postchange patterns are to each other.

We looked at the transformation prediction similarity as a function of initial pattern similarity for 225 pairs of patterns (3 stimuli × 5 change types × 15 runs) for each of the 14 participants. We evaluated this relationship in three ways. First, as a basic measure of the relationship, we calculated the correlation between $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ directly. Next, we used a linear mixed-effects model to predict $r_{transformation\ similarity}$ from $r_{pattern\ similarity}$. This model allowed us to determine how much variance in transformation prediction similarity was predicted by initial pattern similarity. Finally, to ensure that the patterns produced by the transformation analysis indeed better matched the true postchange patterns beyond the initial similarity between the prechange and postchange patterns, we compared the mean $r_{transformation\ similarirty}$ with the mean $r_{pattern\ similarity}$.

Raw image analysis. To test whether the transformations are based on low-level, image-like (arising from retinotopic organization) relationships between input and output patterns, we applied representational transformation analysis to the stimuli directly. We converted the 30 stimuli to grayscale, rescaled them to 10% (51 × 51 pixels), and reshaped them into a 2601-pixel-length one-dimensional vector. We generated 15 "samples" of the same image by adding random Gaussian noise (from −1 to 1) to each so that we could perform the same cross-validation procedure as was performed with the neural data in Experiment 1. There were 450 (30 × 15) images in total. Transformation matrices were computed separately for each object and for each affine change (3 objects × 5 changes). Training was performed on 14 image "samples" and tested on the remaining left-out image. Otherwise, the procedure was identical to the one used for neural patterns. To investigate the effectiveness of the transformations as a function of stimulus integrity, we performed this analysis with six Gaussian noise levels (variance = 0.0, 0.01, 0.05, 0.1, 0.3, and 0.5) added to the images.

Experimental design and statistical analyses. The sample size for Experiment 1 was 14. This size sample was selected to be sufficient for multivoxel pattern detection using brief fMRI acquisition runs (Coutanche and Thompson-Schill, 2012). All statistical tests were within-subject and were conducted using scipy (Jones et al., 2001) or R (R Core Team, 2018).

Patterns were extracted from a large swath of lateral occipital cortex (LOC). We aimed to be as agnostic as possible about the nature of the representations on which the transformations operate. We therefore used a large LOC ROI because it could encompass both retinotopic representations and high-level visual representations. LOC is a high-level object-selective region (Grill-Spector et al., 2001) that codes for some visual features, such as shape (Cant and Goodale, 2007) and size (Konkle and Oliva, 2012; Chiou and Lambon Ralph, 2016). The use of this ROI optimized our ability to find any general transformations and could produce several different interpretable outcomes.

The primary comparison was whether the predicted postchange pattern and the true postchange pattern were the most similar, compared with any of the four control analyses. Pattern similarity was computed using Pearson's correlation. Correlation and correlation distance are among the most reliable similarity metrics for neuroimaging research (Walther et al., 2016). The correlations were always Fisher z-transformed for all analyses. To test for a difference between conditions (e.g., comparing the predicted pattern correlation with the true pattern vs with the wrong pattern), correlations were averaged by condition and by participant and compared using a paired t test. To determine whether change type (e.g., enlarge, rotate+60, etc.) affected the success of the transformations, we averaged correlations by condition, stimulus, and participant and conducted a repeated-measure ANOVA on the predicted pattern correlation with the true pattern versus with the wrong pattern.

To assess the relationship between the initial pattern similarity and the transformation prediction similarity, we calculated the correlation between $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ for each participant and compared the mean with zero using a one-sample t test. Next, to better accommodate the additional factors in our design, we used a linear mixed-effects model (implemented by the *lme4* package in R) (Bates et al., 2015). We used initial similarity as a fixed-factor predictor of transformation similarity, and included stimulus, change type, run, and participant as random intercepts. The model was fit by REML and t tests used the Satterthwaite approximation (implemented by the *lmerTest* package in R) (Kuznetsova et al., 2017) for degrees of freedom and significance level. We used a method of estimating marginal R^2 (i.e., variance explained by the fixed factor alone) outlined in Nakagawa and Schielzeth (2013) and as implemented by the *MuMIn* package in R (Barton, 2018). Finally, we compared mean $r_{transformation\ similarity}$ and mean $r_{pattern\ similarity}$ using a paired t test.

For the image analysis, a repeated-measures ANOVA was run on image similarity (correlation) using analysis (True Change, Wrong Change, Wrong Object) and variance as a predictor, with validation image (rather than participant) treated as a random effect.

Experiment 2: representational transformation analysis over time with MEG

We applied representational transformational analysis to an existing MEG dataset (Isik et al., 2014), which offers higher temporal resolution over fMRI, to determine how early in visual processing these transformations occur. In this new domain, we were able to see the formation of the transformations over time by measuring when in time the transformation matrices generate accurate predictions. Furthermore, we also assessed the stages of the transformation by comparing the efficacy of the transformation matrices when trained and tested at different points in time. The MEG experiment has been described in detail previously (Isik et al., 2014). We therefore only describe the essential features here.

Participants. The original study (Isik et al., 2014) tested 11 participants (3 women) 18 years of age or older with normal or corrected-to-normal vision using a variety of stimuli. From this dataset, we analyzed 8 participants tested with consistent stimuli presentation conditions described below. The Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects approved the experimental protocol. Participants provided informed written consent before the experiment.

MEG recordings and data processing. The MEG scanner used was an Elekta Neuromag Triux with 102 magnetometers at 204 planar gradiometers, and the MEG data were sampled at 1000 Hz. The MEG data were preprocessed using Brainstorm software (Tadel et al., 2011). First, the signals were filtered using Signal Space Projection for movement and sensor contamination (Tesche et al., 1995). The signals were also bandpass filtered from 2100 Hz with a linear phase finite impulse response digital filter to remove external and irrelevant biological noise, and the signals were mirrored to avoid edge effects of bandpass filtering. Each sensor's activity was averaged within each 5 ms time window to yield nonoverlapping bins used in the analysis.

Stimuli. Stimuli were computer-generated images of six objects (basketball, bowling ball, football, adult head, child head, hat). The objects were presented in isolation at three sizes (2°, 4°, and 6° of visual angle in diameter) in the center of the visual field. Only the largest-sized 6° diameter images were shown at three positions (centered, and ±3° vertically). All images were presented in grayscale on a 48 cm × 36 cm display, 140 cm away from the subject; thus, the screen occupied 19° × 14° of visual angle. The largest, central image (6°) was chosen as the "prechange" state of the transformation from which the object could either become "small" (4°) or "tiny" (2°), or move "up" (3° vertically) or "down" (−3° vertically) in the final state of the transformation. Thus, there were four affine change types.

Experimental design and procedure. Images were presented for 48 ms with 704 ms interstimulus interval. Image order was randomized for each experiment, and each stimulus was repeated 50 times. Participants performed a fixation task while viewing images presented two in a row, in which they indicated whether the fixation cross was the same color or different colors when each image appeared. This ensured central fixation and attention.

Channel selection. Because the operations underlying the affine viewpoint changes used in this experiment were hypothesized to be visual in nature, only data from occipital channels were used (72 channels: 24 magnetometers and 48 gradiometers).

MEG transformation analysis. The representational transformation analysis was applied as described above, except as follows. Transformation matrices were now computed separately for each participant and for each time point (5 ms nonoverlapping bins). MEG data were segmented into 5 folds (each containing 10 stimulus trials): four training folds and one left-out, test fold (40 trials in the training data). Each channel composing the prechange pattern was used as an individual predictor of each channel composing the postchange patterns. For all analyses, we applied representational transformation analysis across different stimuli (e.g., transformations were derived using all combinations of objects for training and validation objects).

A linear model was fit with a 72 [channels] × 40 [trials] matrix of predictors, X (and manually fit intercept) to predict a 72 [channels] × 10 [trials] matrix of dependent variables, Y. The solution, $\hat{\beta}$, in this case, is a 72 × 72 transformation matrix. Once this transformation matrix was computed, it was applied to the prechange pattern of the validation stimulus in the remaining fifth testing fold by taking the dot product of the 72 [channels] × 10 [trials] matrix of predictors and the $n \times n$ transformation matrix to yield a 72 [channels] × 10 [trials] predicted postchange pattern. This process was repeated for every time point (from 200 ms before stimulus onset to 500 ms after stimulus onset; 140 time points each).

We also assessed the stages of the transformation by comparing the efficacy of the transformation matrices when trained and tested at different points in time. For example, if the computations occurring early in visual response differed from those occurring later, then an early-response transformation matrix should predict early response patterns but not late-response patterns, and a late-response transformation matrix should predict late response patterns but not early-response patterns. The procedure for this analysis was identical to the transformation analysis described above, except that it was repeated for each time point pairing (each time point served as the training and testing time point; 19,600 pairs total). The analyses were computed on 32-core cluster over several weeks.

Control analyses and permutation testing. We conducted the four control analyses (Wrong Change, Wrong Object, Mismatched Labels, Scrambled Transformation) exactly as described for Experiment 1, except for as follows: In the Mismatched Label and Scrambled Transformation permutation approaches, only 50 permutations were conducted. These analyses were only conducted for transformations that were trained and tested on the same time point.

Transformation similarity versus initial pattern similarity. We compared correlation between prechange versus postchange patterns (i.e., initial pattern similarity, $r_{pattern\ similarity}$) and the correlation between the predicted postchange pattern and the true postchange pattern (i.e., the True Change transformation prediction similarity, $r_{transformation\ similarity}$) in exactly the same way as described for Experiment 1, except for as follows: We used the mean $r_{pattern\ similarity}$ and $r_{transformation\ similarity}$ around the peak correlation (150–200 ms) to compute the initial pattern similarity and transformation prediction similarity. This allowed us to look at the transformation prediction similarity as a function of initial pattern similarity in the same manner as Experiment 1. There were 120 pairs of patterns (6 stimuli × 4 change types × 5 runs) for each of the eight participants.

Experimental design and statistical analyses. The experimental design and statistical analyses were exactly the same as Experiment 1, except for as follows. The sample size for Experiment 2 was 8. This was the sample available to us for analysis (because Experiment 2 was run on an existing dataset), but it is consistent with other MEG studies investigating object recognition. Patterns were extracted all 72 occipital channels. This was to best align with Experiment 1.

The primary comparison was whether the predicted postchange pattern was most similar to the true postchange pattern compared with those for other objects, and how this relationship unfolded in time. Patterns were obtained for each time point, and pattern similarity was computed using Pearson's correlation. The correlations were always Fisher z-transformed for all analyses. To evaluate the difference between the predicted pattern correlation with the true pattern versus with the wrong pattern, correlations were computed separately for each time point. At each time point, correlations were averaged by condition and by participant, and compared using a paired t test (e.g., comparing the predicted pattern correlation with the true pattern vs with the wrong pattern).

For the temporal generalization analysis, paired t tests were used exactly as described above for all pairs of time points. To correct for multiple comparisons on the off-diagonal, significance level was set at $p < 0.005$, and we plotted only clusters with ≥10 adjacent significant time points.

The relationship between transformation prediction similarity as a function of initial pattern similarity was assessed in exactly the same manner as Experiment 1.

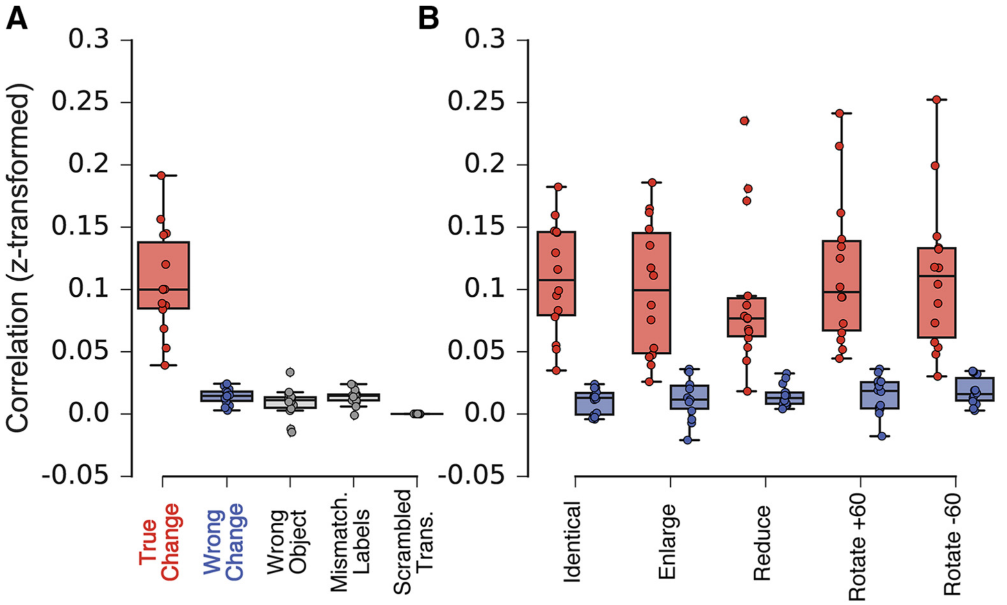

Figure 2. Predictive success of representational transformation analysis. ***A***, The Pearson correlation (Fisher *z*-transformed) between the predicted postchange pattern and the True Change pattern (red), Wrong Change pattern (blue), and Wrong Object pattern. Control analyses' (gray) correlations between the true postchange pattern and the patterns generated by permuting the training labels (Mismatched Labels) and permuting the transformation coefficients (Scrambled Transformation) are also plotted. ***B***, The correlation between the predicted postchange pattern and the True Change pattern (red) and Wrong Change pattern (blue) for each different affine change. Individual points correspond to individual participant means ($n = 14$). Boxes represent quartiles 1–3, and whiskers indicate ±2 SDs.

Results

Experiment 1: representational transformation analysis

We found a main effect of analysis ($F_{(4,52)} = 64.71$, $p = 2 \times 10^{-16}$, $\eta_p^2 = 0.83$, repeated-measured ANOVA) (Fig. 2*A*), wherein the transformation matrices predicted postchange patterns that were more similar to the true postchange pattern compared with any of the control four analyses (p values $<3.26 \times 10^{-6}$, paired t tests). Critically, the predicted pattern was more similar to the true postchange pattern compared with a Wrong Change postchange pattern for the same object. These results indicate that there are transformations that link object representations as the object undergoes a viewpoint change. Furthermore, these transformations can be derived from one object and applied to a novel object to generate a pattern of brain activity that approximates the postchange representation of the novel object. If transformations are mediated simply by the similarity between the prechange and postchange perceptual states, then transformations should not have generalized to novel objects.

Among the control analyses, Scrambled Transformation permutation control analysis produced the smallest correlation between prechange and postchange patterns ($r_{scrambled} = 0.0001 \pm 0.00$), and all of the other control analyses produced correlations that were greater than this baseline ($r_{wrong\ change} = 0.014 \pm 0.006$, $r_{wrong\ object} = 0.009 \pm 0.012$, $r_{mismatched\ labels} = 0.014 \pm 0.006$, p values ≤ 0.0153, Cohen's d values > 0.76, paired t tests), and no significant pairwise differences among these three control analyses (p values > 0.16, Cohen's d values < 0.39, paired t tests). These results suggest that some information may be preserved in these control patterns (compared with the absolute baseline of Scrambled Transformation). For example, object identity information may account for the above-baseline correlations in the Wrong Change and Mismatched Label analyses. But importantly, the correct transformation is far better at approximating the true pattern of activity than any of these sources of information alone.

When we compared the correlations between the predicted postchange pattern for patterns for the True Change and for the Wrong Change across all affine changes, the results were consistent, as is clear from Figure 2*B*. We confirmed this statistically using a repeated-measure ANOVA, which showed that there was no interaction between change type and analysis type ($F_{(4,52)} = 0.64$, $p = 0.64$, $\eta_p^2 = 0.05$). Thus, the transformations were equally effective across all affine changes.

Comparing representational transformation analysis to initial pattern similarity

To further investigate the nature of these transformations, we asked how much the success of our transformations depended on initial pattern similarity between the prechange and true postchange patterns. A similarity-based account of the relationship between prechange and postchange representations predicts that the two states should be highly correlated and, moreover, this correlation should explain the success of the transformations. These types of relationships can appear as representational structure in representational similarity analyses (Kriegeskorte et al., 2008a), whereby classes of objects are represented similarly regardless of changes in viewpoint or exemplar. A high correlation could also be due to limitations of fMRI acquisition: the prechange and postchange patterns were formed by BOLD response measured close in time during the experiment (3966 ms interstimulus interval).

We therefore directly compared transformation prediction similarity (the correlation between the predicted postchange pattern and the true postchange pattern), $r_{transformation\ similarity}$, and the initial pattern similarity, $r_{pattern\ similarity}$. As a rough measure, we first correlated $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ (by correlating these values for each participant and comparing the average correlation to zero with a one-sample t test). We found that the correlation between $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ was significantly above zero ($r = 0.17 \pm 0.15$, $t_{(13)} = 4.22$, $p =$

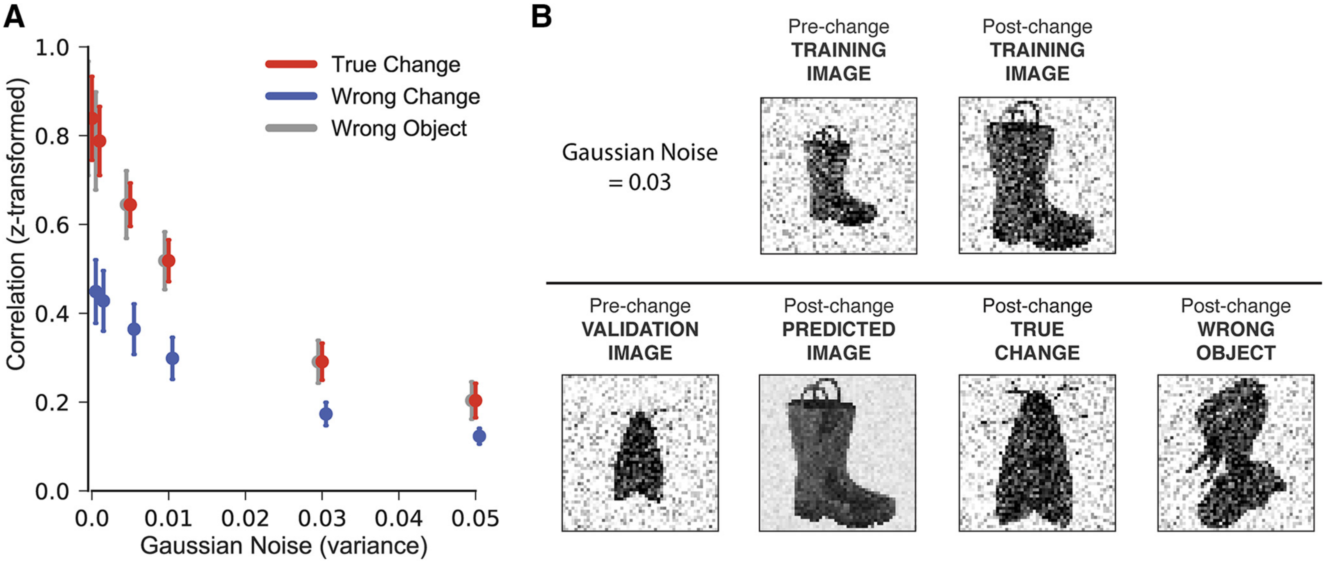

Figure 3. Representational transformation analysis applied to stimulus images directly. Gaussian noise with different variance was added to the images to create multiple samples of the same image for cross-validation. ***A***, The Pearson correlation (Fisher *z*-transformed) between the predicted postchange image and the True Change image (red), Wrong Change image (blue), and Wrong Object image (gray). Points represent the average across stimulus type and affine change type. Error bars indicate 95% CIs. ***B***, An example of the type of image generated by the transformation. The "enlarge" transformation was trained on the manmade stimulus (top row, Gaussian noise with variance = 0.03) and applied to the "natural" validation image. The transformation failed to generalize, as evident by the "enlarged boot" predicted image. This was compared with the True Change and the Wrong Object.

0.001, Cohen's $d = 1.05$). This positive relationship was further confirmed by a linear mixed-effects model, which can better accommodate the other factors (run, stimulus, change type, participant) as random effects. From this, we found that $r_{pattern\ similarity}$ positively predicts $r_{transformation\ similarity}$ ($b = 0.0995$, $t = 8.283$, $p = 2 \times 10^{-16}$), although it only accounted for 2.20% of the variance (R^2). Critically, $r_{transformation\ similarity}$ was greater than $r_{pattern\ similarity}$ ($r_{transformation\ similarity} = 0.11 \pm 0.04$, $r_{pattern\ similarity} = 0.04 \pm 0.08$, $t_{(13)} = 2.43$, $p = 0.0302$, Cohen's $d = 0.65$, paired t test), indicating that applying the transformation to the prechange pattern made it more similar to the true postchange pattern.

Overall, this small but significant relationship between transformation similarity and pattern similarity indicates that perceptual states are neither linked exclusively by similarity-based association nor through transformations. These results show that mere similarity between perceptual states does not explain the success of the transformation analysis and that transformations yield a better approximation of the underlying representation.

Can transformations be learned from raw images?

Finally, an alternative account of our results is that the transformations are simply linking one retinotopic pattern to another because retinotopic organization is known to exist all over LOC. Although we did not measure retinotopic organization directly in this study, one can understand this alternative account by thinking of the activation patterns as "image-like" representations, rather than high-level object representations. Instead of a general, abstract "enlarge" transformation, the transformation may work by simply learning that if voxel n is active in the input pattern, the immediate voxels that surround voxel n should be active in the output pattern. In this case, an input pattern for "moth" would still yield "big moth" and an input pattern for "boot" would still yield "big boot," but the transformation would be based on low-level, image-like (arising from retinotopic organization) relationships between input and output patterns.

A strong test of this alternative account is to determine what happens if we run the same analyses on the raw images themselves: can our representational transformation analysis learn the seemingly simple affine image transformations from the raw images directly, and if so, does such learning generalize to a novel stimulus input? That is, if a representational transformation analysis can learn the relation between the raw images of "moth" and "big moth," can this be applied directly to "boot" to yield an output of "big boot?" If so, it becomes more likely that the transformations are due to image information encoded in image-like representations. If not, it suggests that our analyses learn transformations based on more abstract and high-level representations.

We found a main effect of analysis ($r_{True\ Change} = 0.55 \pm 0.24$, $r_{Wrong\ Change} = 0.31 \pm 0.13$, $r_{Wrong\ Object} = 0.55 \pm 0.25$) ($F_{(2,4)} = 15.89$, $p = 0.0125$, $\eta_p^2 = 0.89$, repeated-measures ANOVA) (Fig. 3*A*, red vs blue). There was a main effect of variance ($F_{(1,2)} = 8393.16$, $p = 0.0001$, $\eta_p^2 = 1.0$, repeated-measures ANOVA). We also found an interaction between analysis and variance ($F_{(2,4)} = 16.46$, $p = 0.0117$, $\eta_p^2 = 0.89$).

Combining 15 samples at each of the 6 variance levels, we replicated the finding that the predicted postchange image best matched true postchange image (True Change) compared with the wrong postchange images (Wrong Change) ($r_{True\ Change} = 0.55 \pm 0.24$, $r_{Wrong\ Change} = 0.31 \pm 0.13$, $t_{(89)} = 17.196$, $p = 4.9 \times 10^{-30}$, Cohen's $d = 1.82$, paired t test) (Fig. 3*A*, red vs blue).

However, in strong contrast to our neural results that showed general transformations that produced predicted patterns that matched both the specific affine change and specific object, the predicted postchange image was just as similar to a different object that had undergone the same change (Wrong Object), averaged across all levels of noise ($r_{True\ Change} = 0.55 \pm 0.24$, $r_{Wrong\ Object} = 0.55 \pm 0.25$, $t_{(89)} = 0.011$, $p = 0.9914$, Cohen's $d < 0.01$) (Fig. 3*A*, red vs gray).

Thus, the image-based transformations were learning something different from the neural transformations. Unlike with neural patterns, the postchange predicted images are readily interpretable (because they should be images of something). Therefore, we investigated the predicted transformed images directly to

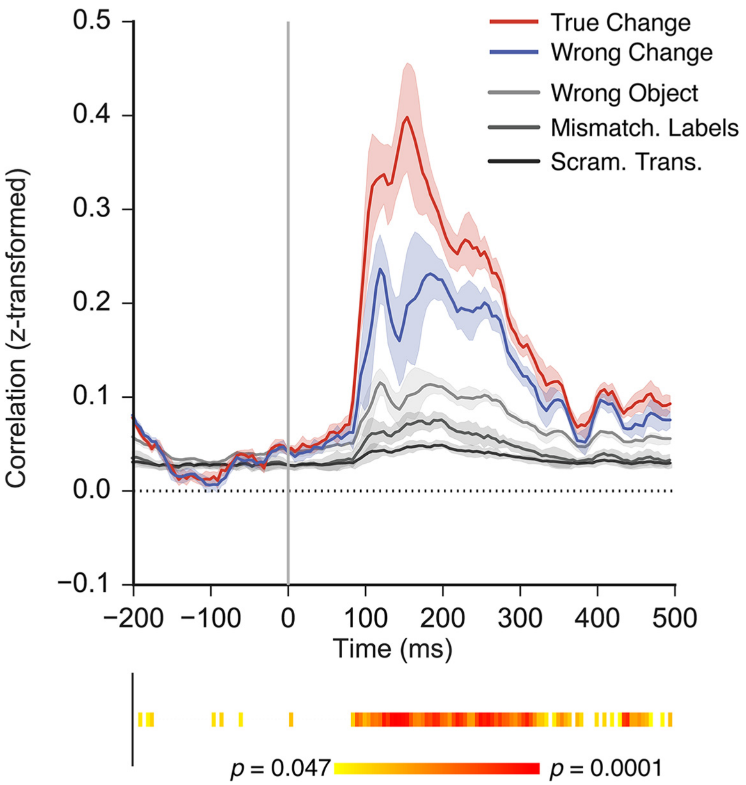

Figure 4. The time course of transformations. The Pearson correlation (Fisher *z*-transformed) between the predicted pattern and True Change pattern (red) and the Wrong Change pattern (blue) is plotted as function of time. Solid gray vertical bar represents the onset of the stimulus. Two-tailed paired *t* tests between True Change and Wrong Change were computed at each time point, and significant differences ($p < 0.05$) between the two conditions are plotted below the graph. Shaded areas for True Change and Wrong Change represent 95% CI of the within-subject mean difference ($n = 8$). Control analyses' correlations between the predicted pattern and Wrong Object pattern (light gray), the true postchange pattern, and the patterns generated by permuting the training labels (Mismatched Labels, gray) and permuting the transformation coefficients (Scrambled Transformation, dark gray) are also plotted. For the control analyses, shaded areas represent 95% CI.

determine whether we could better understand this discrepancy. Instead of learning an abstract affine transformation linking the two training images, the transformation produced the postchange image for whatever stimulus upon which it had been trained. For example, a transformation for "enlarge" trained on the boot stimulus, only produced a "big boot" predicted image, regardless of the validation image to which it was applied (Fig. 3*B*).

Because all stimuli were used as True Change validation stimuli and Wrong Object comparison stimuli, the average correlation for True Change and Wrong Object was basically the same. That is, the correlation between "big boot" and "big moth" would serve as the True Change correlation if "moth" were the validation image and the correlation between "big boot" and "big greeble" would serve as the Wrong Object correlation. But the opposite would be true if "greeble" were the validation image. This resulted in the predicted postchange image being just as similar to the wrong object as the true object. These results show that the regression used in our transformation analysis (L2-regularized linear regression) fails to learn the affine relationship between these 2D images at the pixel level.

Overall, these results show that the transformations between representations of objects that have undergone a common affine change generalize across different objects. This is the first demonstration showing that a transformation that changes existing representations can be isolated; and when applied to a novel representation, the transformation produces an accurate prediction of a new representation. Furthermore, these transformations do not rely on image-like representations alone, and therefore reflect general transformations between high-level object representations.

Experiment 2: the time course of representational transformation

To determine the time course of the transformation, we derived and applied the transformation matrix at the same time point. From ~90 to 300 ms after stimulus onset, and then again after 450 ms, transformation matrices computed from five different stimuli could predict postchange patterns for a new stimulus undergoing the same affine change, compared with the Wrong Change patterns (Fig. 4). The Wrong Change correlations were also above zero (similar to Experiment 1), suggesting that some information may be preserved in these control patterns (e.g., object identity information), but importantly, the correct transformation was far better at approximating the true pattern of activity.

Therefore, using a different imaging modality, different stimuli, and different affine changes, representational transformation analysis yielded results that demonstrate general transformations among different objects that have undergone an affine change. These results also show that the transformations are successful at predicting novel representations very soon after the presentation of an image and then again somewhat later in processing.

Comparing representational transformation analysis with initial pattern similarity

To assess whether the success of our transformations depended on initial pattern similarity between the prechange and true postchange patterns, we computed the mean correlation between prechange versus postchange patterns (i.e., initial pattern similarity, $r_{pattern\ similarity}$) from 150 to 200 ms and compared it with the mean correlation between the predicted postchange pattern and the true postchange pattern (i.e., transformation prediction similarity, $r_{transformation\ similarity}$) from 150 to 200 ms. We looked at prediction similarity as a function of initial pattern similarity for 120 pairs of patterns (6 stimuli × 4 change types × 5 runs) for each of the 8 participants.

First, we first correlated $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ (by correlating these values for each participant and comparing the average correlation to zero with a one-sample *t* test). We found that the correlation between $r_{transformation\ similarity}$ and $r_{pattern\ similarity}$ was significantly above zero ($r = 0.25 \pm 0.23$, $t_{(7)} = 2.911$, $p = 0.0226$, Cohen's $d = 2.26$). This positive relationship was further confirmed by a linear mixed-effects model. From this, we found that $r_{pattern\ similarity}$ positively predicts $r_{transformation\ similarity}$ ($b = 0.3982$, $t = 8.742$, $p = 2.29 \times 10^{-18}$), although it only accounted for 8.26% of the variance (R^2). Critically, $r_{transformation\ similarity}$ was greater than $r_{pattern\ similarity}$

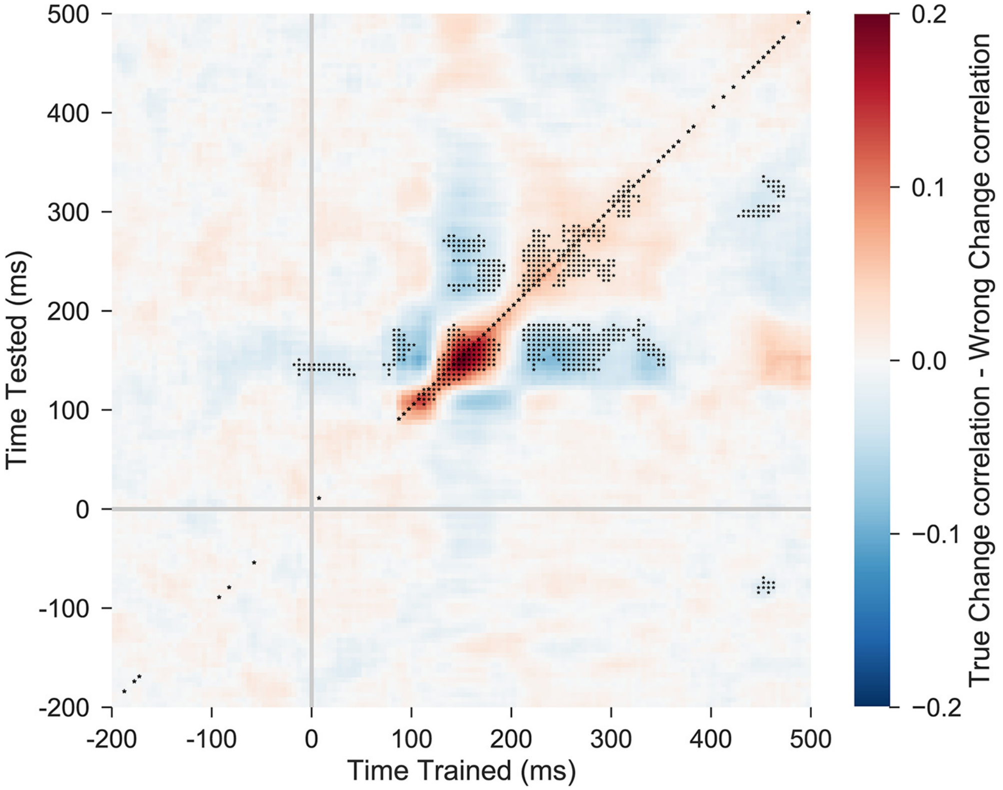

Figure 5. The stages of transformation. Transformations were derived at each training time point (*x*-axis) and applied at each testing time point (*y*-axis). The difference between the True Change correlation values and Wrong Change correlation values is plotted. Darker red represents that the transformation was more successful. Two-tailed, paired *t* tests between True Change and Wrong Change were computed for each training and testing time point. *Significant differences. The diagonal reflects the within-time point significance levels ($p < 0.047$) from Figure 4. To correct for multiple comparisons on the off-diagonal, significance level was set at $p < 0.005$ and only clusters with ≥10 adjacent significant time points are plotted ($n = 8$).

($r_{transformation\ similarity} = 0.34 \pm 0.10$, $r_{pattern\ similarity} = 0.16 \pm 0.06$, $t_{(7)} = 10.98$, $p = 1.15 \times 10^{-5}$, Cohen's $d = 3.88$, paired t test), indicating that applying the transformation to the prechange pattern made it more similar to the postchange pattern.

Overall, across changes in viewpoint, these results show that mere similarity between perceptual states does not explain the success of the transformation analysis.

Stages of transformation

The transformations appeared to produce accurate patterns at two different points in time: an early interval from ~100 ms to 300 ms and a late interval at ~450 ms. Previous studies have found significant classification accuracy at early intervals (Carlson et al., 2013; Isik et al., 2014), with neural decoding ~100 ms correlated with V1 and decoding between 250 and 500 ms correlated with inferior temporal (Cichy et al., 2014). Do these two different intervals correspond to two different computational stages in our study?

To answer this question, we applied representational transformation analysis across time points: the transformation was computed based on training data drawn from one time point and tested on test data from all other time points (including test time points occurring before the training time point). This approach is a type of temporal generalization method (King and Dehaene, 2014), except we computed transformations rather than training a classifier at one time point and applying it at the others. (For object representations, there is very little temporal generalization for invariant representations, e.g., Isik et al., 2014). Critically, can the transformations trained at the early interval predict patterns at the late interval, or vice versa?

Several key results are clear from inspecting Figure 5, which plots the difference between the True and Wrong pattern correlations. For direct comparison, the diagonal reflects the significance levels from Figure 4 (asterisks along the diagonal in Figure 5 denote significance at $p < 0.047$, as shown in the colorbar in Figure 4), but to correct for the multiple comparisons at all other time points, only clusters of 10 or more adjacent, significant ($p < 0.005$) time points are displayed. First, the transformations are most effective from ~120 to 200 ms. Second, there appear to be two distinct stages: first from 120 to 200 ms and from 220 to 320 ms. Finally, a transformation produced at a particular stage is only effective within a narrow interval around the time at which it was generated. For example, the transformations trained at the first stage are effective in the first stage (evidenced by the dark red positive clusters) but produce patterns that are worse than the wrong patterns during the second stage (evidenced by the blue negative clusters directly above the dark red cluster). These results could arise from inhibition of return or an adaptation effect. As the transformation occurs, the intermediate stages may be actively suppressed, so that if one trains on an earlier time point and tests on a later time point (or vice versa), the true postchange pattern at the later time point will contain an attenuated or even opposing signal. This would decrease the similarity between the predicted postchange pattern and the True postchange pattern. At that point, any invariant object information among the patterns may drive the similarity, causing the Wrong patterns to be more similar because they contain information about the same object.

The same pattern of results was found for the second stage: the transformations trained at the second stage are effective during the second stage, and produce worse patterns during the first stage (evidenced by the blue cluster directly below the second stage clusters, and to the right of the first stage cluster).

Although the earliest time points at which the transformation was effective (~90–120 ms; Fig. 4) did not produce a cluster that survived multiple comparisons correction, the results plotted in Figure 5 do suggest that the transformations occurring at the earliest time points may be different from those occurring immediately afterward: when the transformations are trained at these early time points and tested at 150 ms, they produce a significant negative cluster. This on-diagonal pattern of results is consistent with the pattern of results produced by MEG decoding of inferior temporal, whereas decoding of V1 produced off-diagonal temporal generalization (Cichy et al., 2014).

Overall, these results show strong evidence for two distinct stages of transformation, and possibly even a third that occurs at the earliest stage of visual processing.

Discussion

It has long been thought the visual system applies a set of general procedures to visual representations to extract properties and relations not explicit in the visual input itself. These "visual routines" are operations that act on base representations to produce incremental representations (Ullman, 1984). While these operations typically encompass early visual analyses, such as intersection and boundary tracking, our results show that a similar mechanism exists for affine visual changes, such as viewpoint change. While the brain must accommodate "online" changes that occur in real time (e.g., integrating a visual scene across saccades or tracking moving objects), it is also important to link object state representations, such as before and after an object undergoes a change. Using a novel method, representational transformation analysis, our study isolated general transformations, whereby a transformation function can be applied to a brain representation of an object to transform it into the representation of the object after it itself has undergone a change in viewpoint. This shows that operations are preserved across stimuli and that the prechange and postchange representations are not linked simply through similarity.

On the one hand, it is extremely important to recognize objects across such changes. Although we store in memory individual "snapshots" of an object as our view of it changes (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Tarr, 1995), there is extensive evidence of viewpoint-invariant coding of objects across changes in position, size, or viewpoint (DiCarlo and Cox, 2007). This property of neuronal coding increases along the ventral visual pathway (Rust and DiCarlo, 2010), including LOC (e.g., Grill-Spector et al., 1999). On the other hand, it is also extremely important to recognize when an object has changed in some way, such as during a viewpoint change: just as we would never mistake a chair for a different chair as our viewpoint changes, we would also never mistake a side view for top view just because we recognize the chair. Thus, invariant object coding does not obviate the need for coding of specific changes in viewpoints. Our results show that the transformation from one viewpoint to another applies across objects.

In this study, we have focused on affine changes, such as those we experience for certain viewpoint changes. But transformations may be important for changes more generally. A number of recent studies have shown that there is at least a coarse representation of space in LOC (e.g., MacEvoy and Epstein, 2007; Carlson et al., 2011), which is consistent with the view that object representations are selective for both position and objects (Edelman and Intrator, 2000, 2003; Kravitz et al., 2008). However, under this view, there are multiple position-specific representations for each object that do not need to be transformed to be compared. Our results may help reconcile position-dependent and position-independent views by showing that there are transformations specific to a particular position change that act on object representations generally. It has been proposed that the computations that produce invariant object recognition and the computations involved in transforming object representations, such as those demonstrated in this study must operate simultaneously (Cichy et al., 2011). The time course of transformation we found in the current study closely resembles the time course of invariant representation found in Isik et al. (2014), lending some empirical support for this view. Beyond the relevance of transformation for the coding of objects and space, we are also surprisingly good at recognizing when a common transformation has occurred to different objects (Schmidt and Fleming, 2016), such as when a towel or aluminum can have been twisted. Rather than due to high-level inference, our ability to appreciate such changes may arise from visual processing (Chen and Scholl, 2016).

What is the nature of the computations captured by the transformation? It is important to note that our method of finding the transformation (e.g., regularized linear regression) is unlikely to be the method by which the neuronal population achieves the transformation, nor is it a direct algorithmic explanation (per Marr, 1982). The multivoxel representations our transformation analysis relies on are themselves just approximations of the true neural representations, generated by statistical tests (i.e., GLM on fMRI signal), and accordingly, our method of linking the representations can only approximate the operations. Thus, our approach is abstracted from online, neural transformations. However, the strength of our results comes from demonstrating the transformations obtained from one set of representations can be applied to a novel input to generate a new, accurate representation.

Furthermore, the success of our method to learn transformations between neural patterns and the failure of our method to learn transformations between raw images suggest some possibilities about the nature of the computations. How is it that an analysis that fails to capture simple affine changes between images could capture an abstract, general relationship between high-dimensional object representations in the brain?

First, although it is trivial to *apply* an affine change to an image on a computer, it is far from trivial to *learn* a transformation function from prechange and postchange image pairs. While one can easily create a function to rotate an image in-plane, reverse engineering in-plane rotation is different and difficult, even when the problem space is very constrained, such as trying to determine the angle of rotation between two images (there are multiple approaches to solving this, ranging from registering the images through incremental steps until the error between them is minimized, to FFT-based approaches, to feature-detection approaches, either by finding edges and corners or by finding high-level relationships among items/features in the images). Learning functions from raw data alone is extremely complex. In neuroscience, this problem has been approached using encoding models (e.g., Naselaris et al., 2011), in which features of the raw data (usually natural images of scenes) are used to predict the activity of single voxels in the brain. In contrast to the transformations in the present study which link object representations before and after a change, encoding models link brain activity to the stimulus and can therefore be used to reconstruct a stimulus (Nishimoto et al., 2011). However, encoding models are not designed to learn functions that occur in the world from the raw images, only which features reliably cause a certain pattern of activity. To learn functions from raw data alone, pioneers in this domain rely on sophisticated computational search algorithms (which have been used to uncover several laws of physics from scratch, directly from data) (Schmidt and Lipson, 2009) and deep neural networks (which have been shown to learn "intuitive" models of the physical world) (Byravan and Fox, 2016). Therefore, it is unlikely that the computations captured by the transformations operate directly on image-like representations.

Second, the primary function of the brain is to adaptively interact with the world, which requires both representing and transforming incoming information. Therefore, LOC represents information that has already been transformed as it has been processed through the early visual system. Retinotopic representations are extremely high-dimensional, and different points in this high-dimensional space represent each distinct view or encounter of an object. Furthermore, different objects may activate overlapping regions in this space, making the representations extremely intertwined and not linearly separable (DiCarlo and Cox, 2007). The function of the visual system is to change these "tangled" representations into more distinguishable forms, for which a linear decision boundary could be established eventually (e.g., to recognize one object from another). Linear classification of brain patterns has become ubiquitous in cognitive neuroscience (Lewis-Peacock and Norman, 2014), and there is evidence showing that patterns can be combined linearly to predict more sophisticated representations (e.g., in the domain of object and scene recognition) (MacEvoy and Epstein, 2009, 2011) and in the domain of concept formation (Baron et al., 2010; Baron and Osherson, 2011). Just as deep neural networks must transform pixel representations to representational forms that are more general and abstract, the visual system must too accomplish this in some way (Kriegeskorte, 2015; Yamins and DiCarlo, 2016). This visual "preprocessing" of information most likely consists of complex, nonlinear computations, which our method takes advantage of to produce general linear transformations that distinguish both change type and object type. We were successful at isolating the transformations for simple affine changes for a small number of objects, but it may be necessary to use more sophisticated models to investigate other more abstract and fascinating transformations, such as the transformations that give rise to the perception of animacy (Heider and Simmel, 1944) or causal history (Leyton, 1989).

Transformations are not only an integral part of our visual experience, but also of much of cognition. We expect that the representational transformation analysis presented here will be useful to many researchers investigating the representational structures and computational procedures that operate on those structures.

References

Baron SG, Osherson D (2011) Evidence for conceptual combination in the left anterior temporal lobe. Neuroimage 55:1847–1852. CrossRef Medline

Baron SG, Thompson-Schill SL, Weber M, Osherson D (2010) An early stage of conceptual combination: superimposition of constituent concepts in left anterolateral temporal lobe. Cogn Neurosci 1:44 –51. CrossRef Medline

Barton K (2018) MuMIn: multi-model inference. Available at https://CRAN.R-project.org/package=MuMIn.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1– 48.

Bülthoff HH, Edelman S (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc Natl Acad Sci U S A 89:60 – 64. CrossRef Medline

Byravan A, Fox D (2016) SE3-Nets: learning rigid body motion using deep neural networks. Available at http://arxiv.org/abs/1606.02378.

Cant JS, Goodale MA (2007) Attention to form or surface properties modulates different regions of human occipitotemporal cortex. Cereb Cortex 17:713–731. CrossRef Medline

Carlson T, Hogendoorn H, Fonteijn H, Verstraten FA (2011) Spatial coding and invariance in object-selective cortex. Cortex 47:14 –22. CrossRef Medline

Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of object vision: the first 1000 ms. J Vis 13:1. CrossRef Medline

Chen YC, Scholl BJ (2016) The perception of history: seeing causal history in static shapes induces illusory motion perception. Psychol Sci 27:923–930. CrossRef Medline

Chiou R, Lambon Ralph MA (2016) Task-related dynamic division of labor between anterior temporal and lateral occipital cortices in representing object size. J Neurosci 36:4662– 4668. CrossRef Medline

Cichy RM, Chen Y, Haynes JD (2011) Encoding the identity and location of objects in human LOC. Neuroimage 54:2297–2307. CrossRef Medline

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neurosci 17:455– 462. CrossRef Medline

Coutanche MN, Thompson-Schill SL (2012) The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. Neuroimage 61:1113–1119. CrossRef Medline

deCharms RC, Zador A (2000) Neural representation and the cortical code. Annu Rev Neurosci 23:613– 647. CrossRef Medline

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341. CrossRef Medline

Edelman S, Bülthoff HH (1992) Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Res 32:2385–2400. CrossRef Medline

Edelman S, Intrator N (2000) (Coarse coding of shape fragments) + (retinotopy) ≈ representation of structure. Spat Vis 13:255–264. CrossRef Medline

Edelman S, Intrator N (2003) Towards structural systematicity in distributed, statically bound visual representations. Cogn Sci 27:73–109. CrossRef

Fischl B (2012) FreeSurfer. Neuroimage 62:774 –781. CrossRef Medline

Gauthier I, James TW, Curby KM, Tarr MJ (2003) The influence of conceptual knowledge on visual discrimination. Cogn Neuropsychol 20:507–523. CrossRef Medline

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R (1999) Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24:187–203. CrossRef Medline

Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. Vision Res 41:1409 –1422. CrossRef Medline

Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57:243. CrossRef

Hindy NC, Turk-Browne NB (2016) Action-based learning of multistate objects in the medial temporal lobe. Cereb Cortex 26:1853–1865. CrossRef Medline

Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object recognition in the human visual system. J Neurophysiol 111:91–102. CrossRef Medline

Jones E, Oliphant T, Peterson P, others (2001) SciPy: oOpen source scientific tools for Python. Available at http://www.scipy.org/.

Kamitani Y, Sawahata Y (2010) Spatial smoothing hurts localization but not information: pitfalls for brain mappers. Neuroimage 49:1949 –1952. CrossRef Medline

King JR, Dehaene S (2014) Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn Sci 18:203–210. CrossRef Medline

Konkle T, Oliva A (2012) A real-world size organization of object responses in occipitotemporal cortex. Neuron 74:1114 –1124. CrossRef Medline

Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object recognition? Trends Cogn Sci 12:114 –122. CrossRef Medline

Kriegeskorte N (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci 1:417– 446. CrossRef Medline

Kriegeskorte N, Mur M, Bandettini P (2008a) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 2:4. CrossRef Medline

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126 –1141. CrossRef Medline

Kuznetsova A, Brockhoff PB, Christensen RH (2017) lmerTest package: tests in linear mixed-effects models. J Stat Softw 82:1–26.

Lewis-Peacock JA, Norman KA (2014) Multi-voxel pattern analysis of fMRI data. In: The cognitive neurosciences (Gazzaniga M, Mangun R, eds), pp 911–920. Cambridge, MA: Massachusetts Institute of Technology.

Leyton M (1989) Inferring causal history from shape. Cogn Sci 13:357–387. CrossRef

MacEvoy SP, Epstein RA (2007) Position selectivity in scene-and object-responsive occipitotemporal regions. J Neurophysiol 98:2089 –2098. CrossRef Medline

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

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

Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. Cambridge, MA: Massachusetts Institute of Technology.

Nakagawa S, Schielzeth H (2013) A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol 4:133–142. CrossRef

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400 – 410. CrossRef Medline

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646. CrossRef Medline

Peirce JW (2007) PsychoPy, psychophysics software in python. J Neurosci Methods 162:8 –13. CrossRef Medline

R Core Team (2018) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org/.

Rust NC, DiCarlo JJ (2010) Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. J Neurosci 30:12978 –12995. CrossRef Medline

Schmidt F, Fleming RW (2016) Visual perception of complex shape-transforming processes. Cogn Psychol 90:48 –70. CrossRef Medline

Schmidt M, Lipson H (2009) Distilling free-form natural laws from experimental data. Science 324:81– 85. CrossRef Medline

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 [Suppl 1]:S208 –S219.

Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM (2011) Brainstorm: a user-friendly application for MEG/EEG analysis. Comput Intell Neurosci 2011:879716. CrossRef Medline

Tarr MJ (1995) Rotating objects to recognize them: a case study on the role of viewpoint dependency in the recognition of three-dimensional objects. Psychon Bull Rev 2:55– 82. CrossRef Medline

Tesche CD, Uusitalo MA, Ilmoniemi RJ, Huotilainen M, Kajola M, Salonen O (1995) Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. Electroencephalogr Clin Neurophysiol 95:189 –200. CrossRef Medline

Ullman S (1984) Visual routines. Cognition 18:97–159. CrossRef Medline

Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137:188 –200. CrossRef Medline

Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA (2010) Greater neural pattern similarity across repetitions is associated with better memory. Science 330:97–101. CrossRef Medline

Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356 –365. CrossRef Medline