A fast, invariant representation for human action in the visual system

Leyla Isik[*], Andrea Tacchetti[*], and Tomaso Poggio

Center for Brains, Minds, and Machines, MIT

Corresponding author: Leyla Isik
77 Massachusetts Avenue
Bldg 46-4141D
Cambridge, MA 02139
617.258.6933
lisik@mit.edu

[*] These authors contributed equally to this work

Abstract

Humans can effortlessly recognize others' actions in the presence of complex transformations, such as changes in viewpoint. Several studies have located the regions in the brain involved in invariant action recognition, however, the underlying neural computations remain poorly understood. We use magnetoencephalography (MEG) decoding and a dataset of well-controlled, naturalistic videos of five actions (run, walk, jump, eat, drink) performed by different actors at different viewpoints to study the computational steps used to recognize actions across complex transformations. In particular, we ask when the brain discriminates between different actions, and when it does so in a manner that is invariant to changes in 3D viewpoint. We measure the latency difference between invariant and non-invariant action decoding when subjects view full videos as well as form-depleted and motion-depleted stimuli. We were unable to detect a difference in decoding latency or temporal profile between invariant and non-invariant action recognition in full videos. However, when either form or motion information is removed from the stimulus set, we observe a decrease and delay in invariant action decoding. Our results suggest that the brain recognizes actions and builds invariance to complex transformations at the same time, and that both form and motion information are crucial for fast, invariant action recognition.

New and Noteworthy

The human brain can quickly recognize actions despite transformations that change their visual appearance. We use neural timing data to uncover the computations underlying this ability. We find that within 200ms action can be read out of MEG data, and that this representation is invariant to changes in viewpoint. We find form and motion are needed for this fast action decoding, suggesting that the brain quickly integrates complex spatiotemporal features to form invariant action representations.

Keywords: Action recognition, Magnetoencephalography, Neural decoding, Vision

Introduction

As a social species, humans rely on recognizing the actions of others in their everyday lives. We quickly and effortlessly extract action information from rich dynamic stimuli, despite variations in the visual appearance of action sequences, due to transformations such as changes in size, position, actor, and viewpoint (e.g., is this person running or walking towards me, regardless of which direction they are coming from). The ability to recognize actions, the middle ground between action primitives (e.g., raise the left foot and move it forward) and activities (e.g., playing basketball) (Moeslund and Granum 2001), is paramount to humans' social interactions and even survival. The computations driving this process, however, are poorly understood. This lack of computational understanding is evidenced by the fact that even state of the art computer vision algorithms, convolutional neural networks, which match human performance on object recognition tasks (He et al. 2015), still drastically underperform humans on action recognition tasks (Le et al. 2011; Karpathy et al. 2014). In particular, what makes action and other visual recognition problems challenging are transformations (such as changes in scale, position and 3D viewpoint) that alter the visual appearance of actions, but are orthogonal to the recognition task (DiCarlo and Cox 2007).

Several studies have attempted to locate the regions in the brain involved in processing actions, and in some cases, locate regions in the brain containing viewpoint-invariant representations. In humans and nonhuman primates, the extrastriate body area (EBA) has been implicated in recognizing human form and action (Downing et al. 2001; Michels et al. 2005; Lingnau and Downing 2015), and the superior temporal sulcus (STS) has been implicated in recognizing biological motion and action (Perrett et al. 1985; Oram and Perrett 1996; Grossman et al. 2000; Vaina et al. 2001; Grossman and Blake 2002; Beauchamp et al. 2003; Peelen and Downing 2005; Vangeneugden et al. 2009). The posterior portion of the STS (pSTS) represents particular types of biological motion data in a viewpoint invariant manner (Grossman et al. 2010;

Vangeneugden et al. 2014). Beyond visual cortex, action representations have been found in human parietal and premotor cortex when people perform and view certain actions, particularly hand grasping and goal-directed behavior (analogous to monkey “mirror neuron” system) (Hamilton and Grafton 2006; Dinstein, Gardner, et al. 2008; Dinstein, Thomas, et al. 2008; Oosterhof et al. 2010, 2012, 2013; Freeman et al. 2013). However, recent work suggests that these “mirror neuron” regions do not code the abstract, invariant representations of actions, which are coded in visual regions (Wurm et al. 2015, 2016).

Here we investigate the neural dynamics of action processing, rather than the particular brain regions involved, in order to elucidate the underlying computations. We use magnetoencephalography (MEG) decoding to understand when action information is present and how the brain computes representations that are invariant to complex, non-affine transformations such as changes in viewpoint. Timing information can constrain the computations underlying visual recognition by informing when different visual representations are computed. For example, recent successes in MEG decoding have revealed interesting properties about invariant object recognition in humans, mainly that it is fast and highly dynamic, and that varying levels of abstract categorization and invariance increase over the first 200ms following image onset (Carlson et al. 2011, 2013; Cichy et al. 2014; Isik et al. 2014).

Prior work has shown that biological motion can be distinguished from spatially scrambled dots (Hirai et al. 2003; Hirai and Hiraki 2006; Pavlova et al. 2007) and inverted figures (Jokisch et al. 2005) within 200 ms. However, it remains unknown when neural signals can not only detect, but discriminate between different types of biological motion. We use timing data to ask first, when the brain can discriminate between different actions, and second, when it computes invariance to complex, non-affine transformations. Previous studies of invariant recognition of static faces and objects suggest that 3D-viewpoint invariance develops at later stages in the visual processing hierarchy (Logothetis and Sheinberg 1996; Freiwald and Tsao 2010; Leibo et al. 2017). Does this hold for invariant action recognition?

Our results show that we can read out actions as early as 200 ms after a video begins. We further find that the MEG signals are already invariant to changes in viewpoint, suggesting that the brain performs both action recognition and invariance at the same processing stage. We further show that two types of action information, form (as tested with static images) and motion (as tested with point light figures), both contribute to these immediately view-invariant representations. When either form or motion information is removed, view-invariant decoding is lower accuracy and delayed. These results suggest that features that are rich in form and motion content drive the fast, invariant representation of the actions in the human brain.

Materials and Methods

Action recognition dataset

To study the effect of changes in view on action recognition, we used a dataset of five actors performing five different actions (drink, eat, jump, run and walk) on a treadmill from two different views (0 and 90 degrees from the front of the actor/treadmill; the treadmill rather than the camera was rotated in place to film from different viewpoints) [Figure 1] (Tacchetti et al. 2016). These actions were selected to be highly familiar, and thus something subjects would have experienced under many viewing conditions, to include both reaching-oriented (eat and drink) and leg-oriented (jump, run, walk) actions, as well as to span both coarse (eat and drink versus run and walk) and fine (eat versus drink and run versus walk) action distinctions. Every video was filmed on the same background, and the same objects were present in each video, regardless of action (e.g., to avoid confounds such as “run” being detected based on the presence of a treadmill and “drink” being detected based on the presence of a water bottle). Each action-actor-view combination was filmed for at least 52-seconds. The videos were then cut into two-second clips that each included at least one cycle of each action, and started at random points in the cycle (for example, a jump may start midair or on the ground). This dataset allows testing of actor and view invariant action recognition, with few low-level confounds.

To explore the roles of form and motion in invariant action representations, we extended this video dataset with two additional components: a form only dataset, consisting of representative single frames for each action, and a motion-only dataset, consisting of point light figures performing the same actions. For the form dataset, the authors selected one frame per video making sure that the selected frames were unambiguous for action identity (special attention was paid to the actions eat and drink to ensure the food or drink was near the mouth, and occluded views to ensure there was some visual information about action). For the motion point light dataset, the videos were put on Amazon Mechanical Turk and workers were asked to label 15 landmarks in every single frame: center of head, shoulders, elbows, hands, torso, hips, knees, and ankles. Three workers labeled each video frame. We used the spatial median of the three independent labels for each frame and landmark to increase the signal to noise ratio, and independently low-pass filtered the time series (Gaussian Filter with a 30 frames aperture and normalized convolution) for each of the 15 points to reduce the high frequency artifacts introduced by single-frame labeling.

Participants

Three separate MEG experiments were conducted (see below). Ten subjects (5 female, 8 right-handed, age: mean±SD = 28.6±6.1) participated in experiment one, ten subjects (7 female, 10 right-handed, age mean±SD = 25.2±5.0) participated experiment two, and ten subjects (7 female, 9 right-handed, age: mean±SD = 28.3±5.7) participated in experiment three. All subjects had normal or corrected to normal vision. The MIT Committee on the Use of Humans as

Experimental Subjects approved the experimental protocol. Subjects provided informed written consent before the experiment.

Experimental procedure

In the first experiment, we assessed if we could read out different actions both within viewpoint (training and testing on videos at 0 degrees or 90 degrees, without any generalization) and across viewpoint, by training and testing on two different views (0 and 90 degrees). In this experiment ten subjects were shown 50 two-second video clips (one for each of five actors, actions, and two views, 0 and 90 degrees), each presented 20 times.

To examine whether form and motion information were necessary to construct invariant action representations, in the second and third experiments we showed subjects limited "form" (static image) or "motion" (point-light walkers) datasets. Specifically, in the second experiment, ten subjects were shown 50 static images (one for each of five actors, actions, and two views, 0 and 90 degrees), which were single frames from the videos in Experiment 1, for 2 seconds presented 20 times each. In the third experiment, ten subjects were shown 10 two-second video clips, which consisted of point-light walkers traced along one actor's videos from two views in experiment one (labelled by Mechanical Turk workers as described above), presented 100 times each.

In each experiment, subjects performed an action recognition task, where they were asked after a random subset of videos or images (in a randomly interspersed 10% of the trials for each video or image) what action was portrayed in the previous image or video. The purpose of this behavioral task was to ensure subjects were attentive and assess behavioral performance on the various datasets. The button order for each action was randomized across trials to avoid systematic motor confounds in the decoding. Subjects were instructed to fixate centrally. The videos were presented using Psychtoolbox to ensure accurate timing of stimulus onset. Each video had a duration of 2s and a 2s inter-stimulus interval. The videos were shown in grayscale at 3 x 5.4 degrees of visual angle on a projector with a 48 cm × 36 cm display, 140 cm away from the subject.

MEG data acquisition and preprocessing

The MEG data were collected using an Elekta Neuromag Triux scanner with 306 sensors, 102 magnetometers at 204 planar gradiometers, and were sampled at 1000 Hz. First the signals were filtered using temporal Signal Space Separation (tSSS) with Elekta Neuromag software. Next, Signal Space Projection (SSP) (Tesche et al. 1995) was applied to correct for movement and sensor contamination. The MEG data were divided into epochs from -500 - 3500 ms, relative to video onset, with the mean baseline activity removed from each epoch. The signals were band-pass filtered from 0.1–100 Hz to remove external and irrelevant biological noise (Acunzo et al. 2012; Rousselet 2012). The convolution between signals and bandpass filter was implemented by wrapping signals in a way that may introduce edge effects at the beginning and end of each trial. We mitigated this issue by using a large epoch window (-500-3500 ms) and testing significance in a manner that takes into account temporal biases in the data (see significance testing below). The above pre-processing steps were all implemented using the Brainstorm software (Tadel et al. 2011).

General MEG decoding methods

MEG decoding analyses were performed with the Neural Decoding Toolbox (Meyers 2013), a Matlab package implementing neural population decoding methods. In this decoding procedure, a pattern classifier was trained to associate the patterns of MEG data with the identity of the action in the presented image or video. The stimulus information in the MEG signal was evaluated by testing the accuracy of the classifier on a separate set of test data. This procedure was conducted separately for each subject and multiple re-splits of the data into training and test data were utilized.

The time series data of the magnetic field measured in each sensor (including both the magnetometers and gradiometers) were used as classifier features. We averaged the data in each sensor into 100 ms overlapping bins with a 10 ms step size, and performed decoding independently at each time point. Decoding analysis was performed using cross validation, where the dataset was randomly divided into five cross validation splits. The classifier was then trained on data from four splits (80% of the data), and tested on the fifth, held out split (20% of the data) to assess the classifier's decoding accuracy.

Decoding - feature pre-processing

To improve signal to noise, we averaged together the ten different trials for each semantic class (e.g. videos of run) in each given cross validation split of each subject's data so there was one data point per stimulus per cross validation split. We next Z-score normalized that data by calculating the mean and variance for each sensor using only the training data. We then performed sensor selection using only the training data, by applying a five-way ANOVA to each sensor's training data to test if the sensor was selective for the different actions. We use sensors that were selective for action identity, i.e., show a significantly greater variation across class than within class, with $p<0.05$ significance based on a F-test (if no sensors were deemed significant, the one with the lowest p-value is selected). The selected sensors were then fixed and used for testing. To avoid circularity in our feature pre-processing, the test data was never used for the z-scoring or feature selection.

Each sensor (including both magnetometers and gradiometers) was considered as an independent sensor input into this algorithm, and the feature selection, like the other decoding steps is performed separately at each 100ms time bin, and thus a different number of sensors was selected for each subject at each time bin. The average number of sensors selected for each subject across all significant decoding time bins is shown in Table 1. These pre-processing parameters have been shown to empirically improve MEG decoding signal to noise in a previous MEG decoding study (Isik et al. 2014), however as we did not use absolute decoding performance (rather significantly above chance decoding) as a metric for when information is present in the MEG signals, we did not further optimize decoding performance with the present data.

Decoding - classification

The pre-processed MEG data was then input into the classifier. Decoding analyses were performed using a maximum correlation coefficient classifier, which computed the correlation between each test vector and a mean training vector that is created from taking the mean of the training data from a given class. Each test point was assigned the label of the class of the training data with which it was maximally correlated. When we refer to classifier "training" this could alternatively be thought of as learning to discriminate patterns of electrode activity between the different classes in the training data, rather than a more involved training procedure with a more complex classifier. We intentionally chose a very simple algorithm to see in the simplest terms what information is coded in the MEG data. Prior work has also shown empirically that results with a correlation coefficient classifier are very similar to standard linear classifiers like support vector machines (SVMs) or regularized least squares (RLS) (Isik et al. 2014).

We repeated the above decoding procedure at each time bin to assess the decoding accuracy versus time. We re-ran the above procedure 50 times for each subject. We measured decoding accuracy as the average percent correct of the test set data across all decoding runs, and reported decoding results for the average of ten subjects in each experiment. Plots and latency measures were centered at the median value of each of the 100ms time bins.

For more details on these decoding methods see (Isik et al. 2014).

Decoding invariant information

To see if information in the MEG signals could generalize across a given transformation, we trained the classifier on data from subjects viewing the stimuli under one condition (e.g. 0-degree view) and tested the classifier on data from subjects viewing the stimuli under a separate, held out condition (e.g. 90-degree view). This provided a strong test of invariance to a given transformation. In all three experiments, we compared the within and across view decoding. For the "within" view case, the classifier was trained on 80% of data from one view, and tested on the remaining 20% of data from the same view. For the "across" view case, the classifier was trained on 80% of data from one view, and tested on 20% of data from the opposite view, so the same amount of training and test data was evaluated in each case.

Significance testing

We assessed action decoding significance using a permutation test. We ran the decoding analysis for each subject with the labels randomly shuffled to create a null distribution. Shuffling the labels breaks the relationship between the experimental conditions that occurred. We repeated the procedure of shuffling the labels and running the decoding analysis 1000 times to create a null distribution, and reported p-values as the percentage rank of the actual decoding performance within the null distribution.

For each experiment and decoding condition, we averaged the null decoding data across ten subjects and determined when the mean decoding across subjects was above the mean null distribution. We define the decoding "onset time" as the first time the subject-averaged decoding accuracy was greater than the subject-averaged null distribution, with p < 0.05. This provided a measure of when significant decodable information was first present in the MEG signals, and is a standard metric to compare latencies between different conditions (Isik et al. 2013; Cichy et al. 2016). Time of peak decoding accuracy for each condition, an alternative established measure of decoding latency, was found to be much more variable (with 95% confidence intervals that were on average over 400 ms larger than onset times), we therefore restricted ourselves to using onset latency only.

Assessing latency differences

To compare when information arises in different decoding conditions (e.g. within versus across view), we compared onset latency rather than raw decoding performance, because 1) the raw magnitude of a classifier is difficult to interpret 2) we want to know *when* significant information is present in each signal. To compare onset latencies for the within view versus across view decoding, we performed 1000 bootstrap resamples of subjects and use the resulting distribution to compute empirical 95%-confidence intervals (CI) for the onset latency of each condition to estimate the temporal sensitivity of our measure (Hoenig and Heisey 2001), as well as for the difference in onset latency between the two conditions. Specifically, in each bootstrap run, we randomly selected a different subset of ten subjects with replacement, computed onset latencies for each condition (as outlined above) and calculated the difference in onset latency between the invariant and non-invariant conditions. We defined the onset latencies for invariant and non-invariant decoding significantly different with p<0.05 if the empirical 95% interval for their difference did not include 0 (Cichy et al. 2016).

Temporal Cross Training

Beyond decoding latency, we sought to examine the dynamics of the MEG decoding using temporal-cross-training analysis (Meyers et al. 2008; Meyers 2013; Isik et al. 2014; King and Dehaene 2014). In this analysis, rather than training and testing the classifier on the same time point, a classifier was trained with data from one time point and then tested on data from all other time points. Otherwise the decoding methods (including feature pre-processing, cross validation and classification) were identical to the procedure outlined above. This method yielded a matrix of decoding accuracies for each training and test time point, where the rows of the matrix indicate the times when the classifier was trained, and the columns indicate the times when the classifier was tested. The diagonal entries of this matrix contained the results from when the classifier was trained and tested on data from the same time point (identical to the procedure described above).

Results

Readout of actions from MEG data is early and invariant

Ten subjects viewed 2-second videos of five actions performed by five actors at two views (0 degrees and 90 degrees) (Figure 1, top row) while their neural activity was recorded in the MEG. We then trained our decoding classifier on only on one view (0 degrees or 90 degrees), and tested it on the second view (0 degrees or 90 degrees). We could read out action from the subjects' MEG data in the case without any invariance ("within view" condition) at, on average, 250 ms (210-330 ms) (mean decoding onset latency across subjects based on p<0.05 permutation test, 95% confidence intervals of onset latencies reported throughout in parentheses, see Methods) post video onset (Figure 2a, blue trace). Each video began at a random point in a given action sequence, suggesting that the brain can compute this representation from different partial sequences of each action. We also observed a significant rise in decoding after the video offset, consistent with offset responses that have been observed in MEG decoding of static images (Carlson et al. 2011).

We next assessed if the MEG signals were invariant to changes in viewpoint by training the classifier on data from subjects viewing actions performed at one view and testing on a second held out view. This invariant "across-view" decoding arose on average at 230 ms (220-270ms) (Figure 2a, red trace). The within and across view decoding were largely overlapping (Figure 2a, insert), and their onset latencies were not significantly different (p = 0.13), suggesting that the early action recognition signals are immediately view invariant. To ensure that the lack of latency difference between the within and between view conditions was not due to the fact that we are using 100ms overlapping time bins, we re-ran the decoding 10ms time bins and 10ms step size (non-overlapping time bins). Although the overall decoding accuracy was lower, the within and across view decoding onsets were still not significantly different (p = 0.62, Figure 2b).

We next examined which types of actions are decoding in both the within and across decoding conditions. By analyzing the confusion matrices for the within- and across-view decoding, we found that not only are coarse action distinctions made (e.g., between run/walk and eat/drink), but so are fine action distinctions (e.g., between eat and drink) even at the earliest decoding of 250 ms (Figure 3). Further, actions performed in a familiar context (i.e. run and walk on a treadmill) were not better classified than those performed in an unfamiliar context (i.e. eat and drink on a treadmill).

The dynamics of invariant action recognition

Given that the within- and across-view action decoding conditions had similar onset latencies, we further compared the temporal profiles of the two conditions by asking if the neural codes for each condition were stable over time. To test this, we trained our classifier with data at one time point, and tested the classifier at all other time points. This yielded a matrix of decoding accuracies for different train times by test times, referred to as a temporal cross training (TCT) matrix (Meyers et al. 2008; Carlson et al. 2013; Meyers 2013; Isik et al. 2014). The diagonal of this matrix shows when the classifier is trained and tested with data at the same time point, just as the line plots in Figure 2a.

The within-view and across-view TCTs showed that the representations for actions, both with and without view, are highly dynamic as there is little off-diagonal decoding that is significantly above chance (Figure 4a-b). The window of significantly above chance decoding performance from 200-400 ms, in particular, is highly dynamic and decoding only within a 50-100 ms window is significantly above chance. At later time points, the above chance decoding extends to a larger window that spans 300ms, suggesting the late representations for action are more stable across time than the early representations. Further, we find that significant decoding for the within and across view conditions were largely overlapping (Figure 4c) showing that information for both conditions are represented at the same time scale in the MEG data.

Invariant action recognition is impaired in form- and motion-depleted stimuli

To study the roles of two information streams, form and motion, in action recognition, subjects viewed two limited stimulus-sets in the MEG. The first ‘Form’ stimulus set consisted of one static frame from each video (containing no motion information). The second ‘Motion’ stimulus set, consisted of point light figures that are comprised of dots on each actor’s head, arm joints, torso, and leg joints and move with the actor’s joints (containing limited form information) (Johansson 1973). Ten subjects viewed each of the form and motion datasets in the MEG. We could decode action from both datasets in the within view case without any invariance (Figure 5). The early view-invariant decoding that was observed with full movies, however, was impaired for both the form or motion datasets. In the form-only experiment, within view could be read out at 410 ms

(320- ms) and across view at 510ms (430- ms). The onset latencies of 410 ms and 510 ms are the first significantly above chance time points for the average decoding across all ten subjects. Although the average decoding across all ten subjects was significantly above chance, in more than 5% of bootstrap runs (each randomly selecting a *different* subset of ten subjects with replacement, see Methods), the decoding was not significantly above chance. Since we could not calculate a significant onset time in the bootsrap runs that did not reach significantly above chance decoding, the upper limit of the 95% CI for both the within and across view decoding is missing and we did not detect a significant difference between the two conditions. In the motion-only experiment, within view action information could be read out significantly earlier than across view information: 210 ms (180-260 ms) versus 300 ms (300-510 ms), and was significantly different between the two conditions (p = 0.013).

Discussion

We investigated the dynamics of invariant action recognition in the human brain and found that action can be decoded from MEG signals as early as 200 ms post video onset, considerably less than the 2s duration of each video and most action cycles (e.g., one drink from a water bottle). This latency is similar to that found for biological motion detection in evoked responses (Hirai et al. 2003; Jokisch et al. 2005; Hirai and Hiraki 2006; Pavlova et al. 2007). These results are also consistent with a recent MEG decoding study that classified two actions, reaching and grasping, slightly after 200ms post video onset (Tucciarelli et al. 2015). Crucially, we showed that these early neural signals are selective to a variety of full-body actions as well as invariant to changes in 3-D viewpoint.

Interestingly we do not observe a difference in onset latency between invariant and non-invariant action representations. While we cannot completely rule out differences at a finer scale than we can resolve with our methods, this appears to be different than object recognition.

Invariant object information increases along subsequent layers of the ventral stream (Logothetis and Sheinberg 1996; Rust and Dicarlo 2010) causing a delay in invariant decoding relative to non-invariant decoding (Isik et al. 2014). Further, physiology data (Freiwald and Tsao 2010) and computational models (Leibo et al. 2017) of static face recognition have shown that invariance to 3D viewpoint, in particular, arises at a later processing stage than initial face recognition. One possible account of this discrepancy is that even non-invariant ("within view") action representations rely on higher-level visual features (that carry some degree of viewpoint invariant information) than those used in basic object representations.

We characterized the dynamics of action representations using temporal cross training and found that the decoding windows for within and across view decoding are largely overlapping (Figure 4c), suggesting that the beyond onset latencies, the overall dynamics of decoding are similar for non-invariant and view-invariant action representations. It has been suggested that visual recognition, as studied with static object recognition, has a canonical temporal representation that is demonstrated by highly diagonal TCT matrices (King and Dehaene 2014). Our action results generally follow this pattern (Figure 4), but they are more stable over time than previously reported for object decoding (Carlson et al. 2013a; Cichy et al. 2014; Isik et al. 2014).

As shown previously, we find that people can recognize and neural signals can distinguish actions with either biological motion or form information removed from the stimulus (Johansson 1973; Schindler and van Gool 2008; Singer and Sheinberg 2010). In particular, decoding actions within-view is largely intact when form or motion cues are removed. This is likely due to the fact that within-view decoding, unlike the across-view condition, requires little generalization and can thus be performed using low-level cues in the form or motion stimuli. The across-view decoding, on the other hand, requires substantially more generalization and cannot be performed as well, or as quickly as the within-view decoding with form or motion depleted stimuli. It is important to note, however, that the three experiments were completed separately with different subjects, and therefore we cannot directly compare decoding with full videos to the performance with form= or motion-depleted stimuli. Further, while our datasets are a best attempt to isolate form and motion information, it is important to note that static images contain implied motion and that point light figures contain some form information and have less motion information than full movies. Nevertheless, the low-accuracy and delayed invariant decoding with either limited stimulus set suggest that both form and motion information are necessary to build a robust action representation.

Importantly these invariant action representations cannot be explained by low-level stimulus features, such as motion energy as the output of a standard motion energy model (Simoncelli and Heeger 1998) cannot significantly discriminate action across viewpoint (Tacchetti et al. 2016). While we cannot fully rule out the effects of eye movements or shifts in covert attention, eye movement patterns cannot be accounting for our early MEG decoding accuracy, because we do not observe a significant shift in the eye positions between different actions until after 600 ms post video onset and further the same decoder applied to MEG signals does not successfully decode action information using raw eye position data (Figure 2c).

The five actions tested in this study comprise only a small subset of the wide variety of familiar actions we recognize in our daily lives. The five-way classification shows similar decoding across between all five actions, including both coarse and fine action distinctions (Figure 3a-d). These five actions were selected to be highly familiar, and thus we do not know to what extent familiarity is necessary for the immediate invariance we observed. Indeed, modeling and theoretical work suggest that in order to build templates to be invariant to non-affine transformations such as changes in 3-D viewpoint, one must learn templates from different views of each given category (Leibo et al. 2015). It remains an open question how this invariance would translate to unfamiliar actions and how many examples would be needed to learn invariant representations of new actions.

Finally, the longer latency and greater cross-temporal stability of action decoding raises the question of whether recurrent and feedback connections are used to form invariant action representations. This is difficult to test explicitly without high spatiotemporal resolution data. It is indeed likely that feedback and recurrent connections occur within the 200 ms of our earliest decoding (Lamme and Roelfsema 2000). However, further studies have shown that purely feedforward computational models can discriminate actions invariant to viewpoint, and produce representations that explain a significant amount of variance in the human MEG data (Tacchetti et al. 2016).

Taken as a whole, our results show that the brain computes action selective representations remarkably quickly and, unlike in the recognition of static faces and objects, at the same time that it computes invariance to non-affine transformations that are orthogonal to the recognition task. This may represent a key difference between action and object visual processing. Moreover, our findings suggest that both form and motion information are necessary to construct these fast invariant representations of human action sequences. The methods and results presented here provide a framework to study the dynamic neural representations evoked by natural videos, and open the door to probing neural representations for higher level visual and social information conveyed by video stimuli.

Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We thank the McGovern Institute for Brain Research at MIT and the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT for supporting this research. We would like to thank Patrick Winston, Gabriel Kreiman, Martin Giese, Charles Jennings, Heuihan Jhuang, and Cheson Tan for their feedback on this work.

References

Acunzo DJ, Mackenzie G, van Rossum MCW. 2012. Systematic biases in early ERP and ERF components as a result of high-pass filtering. J Neurosci Methods. 209:212–218.

Beauchamp MS, Lee KE, Haxby J V, Martin A. 2003. FMRI responses to video and point-light displays of moving humans and manipulable objects. J Cogn Neurosci. 15:991–1001.

Carlson TA, Hogendoorn H, Kanai R, Mesik J, Turret J. 2011. High temporal resolution decoding of object position and category. J Vis. 11.

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

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

Cichy RM, Pantazis D, Oliva A. 2016. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. Cereb Cortex. bhw135.

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

Dinstein I, Gardner JL, Jazayeri M, Heeger DJ. 2008. Executed and Observed Movements Have Different Distributed Representations in Human aIPS. J Neurosci. 28:11231–11239.

Dinstein I, Thomas C, Behrmann M, Heeger DJ. 2008. A mirror up to nature. Curr Biol. 18:R13-8.

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

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. 2013. A functional and perceptual signature of the second visual area in primates. Nat Neurosci. 16:974–981.

Freiwald WA, Tsao DY. 2010. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science. 330:845–851.

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R. 2000. Brain Areas Involved in Perception of Biological Motion. J Cogn Neurosci. 12:711–720.

Grossman ED, Blake R. 2002. Brain Areas Active during Visual Perception of Biological Motion. Neuron. 35:1167–1175.

Grossman ED, Jardine NL, Pyles JA. 2010. fMR-Adaptation Reveals Invariant Coding of Biological Motion on the Human STS. Front Hum Neurosci. 4:15.

Hamilton AF d. C, Grafton ST. 2006. Goal Representation in Human Anterior Intraparietal Sulcus. J Neurosci. 26:1133–1137.

He K, Zhang X, Ren S, Sun J. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

Hirai M, Fukushima H, Hiraki. 2003. An event-related potentials study of biological motion perception in humans. Neurosci Lett. 344:41–44.

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.

Hoenig JM, Heisey DM. 2001. The Abuse of Power. Am Stat. 55:19–24.

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.

Isik L, Meyers EM, Leibo JZ, Poggio TA. 2013. The dynamics of invariant object recognition in the human visual system. J Neurophysiol. in press.

Johansson G. 1973. Visual perception of biological motion and a model for its analysis. Percept Psychophys. 14:201–211.

Jokisch D, Daum I, Suchan B, Troje NF. 2005. Structural encoding and recognition of biological motion: evidence from event-related potentials and source analysis. Behav Brain Res. 157:195–204.

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. p. 1725–1732.

King J-R, Dehaene S. 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn Sci. 18:203–210.

Lamme VAF, Roelfsema PR. 2000. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 23:571–579.

Le Q V., Zou WY, Yeung SY, Ng AY. 2011. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011. IEEE. p. 3361–3368.

Leibo JZ, Liao Q, Anselmi F, Freiwald WA, Poggio T. 2017. View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation. Curr Biol. 27:62–67.

Leibo JZ, Liao Q, Anselmi F, Poggio TA. 2015. The Invariance Hypothesis Implies Domain-Specific Regions in Visual Cortex. PLOS Comput Biol. 11:e1004390.

Lingnau A, Downing PE. 2015. The lateral occipitotemporal cortex in action. Trends Cogn Sci.

Logothetis NK, Sheinberg DL. 1996. Visual object recognition. Annu Rev Neurosci. 19:577–621.

Meyers EM. 2013. The neural decoding toolbox. Front Neuroinform. 7.

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T. 2008. Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol. 100:1407–1419.

Michels L, Lappe M, Vaina LM. 2005. Visual areas involved in the perception of human movement from dynamic form analysis. Neuroreport. 16:1037–1041.

Moeslund TB, Granum E. 2001. A Survey of Computer Vision-Based Human Motion Capture. Comput Vis Image Underst. 81:231–268.

Oosterhof NN, Tipper SP, Downing PE. 2012. Viewpoint (in)dependence of action representations: an MVPA study. J Cogn Neurosci. 24:975–989.

Oosterhof NN, Tipper SP, Downing PE. 2013. Crossmodal and action-specific: neuroimaging the human mirror neuron system. Trends Cogn Sci. 17:311–318.

Oosterhof NN, Wiggett AJ, Diedrichsen J, Tipper SP, Downing PE. 2010. Surface-Based Information Mapping Reveals Crossmodal Vision–Action Representations in Human Parietal and Occipitotemporal Cortex. J Neurophysiol. 104.

Oram MW, Perrett DI. 1996. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J Neurophysiol. 76:109–129.

Pavlova M, Lutzenberger W, Sokolov AN, Birbaumer N, Krägeloh-Mann I. 2007. Oscillatory MEG response to human locomotion is modulated by periventricular lesions. Neuroimage. 35:1256–1263.

Peelen M V, Downing PE. 2005. Selectivity for the human body in the fusiform gyrus. J Neurophysiol. 93:603–608.

Perrett DI, Smith PAJ, Mistlin AJ, Chitty AJ, Head AS, Potter DD, Broennimann R, Milner AD, Jeeves MA. 1985. Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. Behav Brain Res. 16:153–170.

Rousselet GA. 2012. Does Filtering Preclude Us from Studying ERP Time-Courses? Front Psychol. 3:131.

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.

Schindler K, van Gool L. 2008. Action snippets: How many frames does human action recognition require? In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. p. 1–8.

Simoncelli EP, Heeger DJ. 1998. A model of neuronal responses in visual area MT. Vision Res. 38:743–761.

Singer JM, Sheinberg DL. 2010. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J Neurosci. 30:3133–3145.

Tacchetti A, Isik L, Poggio T. 2016. Spatio-temporal convolutional neural networks explain human neural representations of action recognition.

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.

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.

Tucciarelli R, Turella L, Oosterhof NN, Weisz N, Lingnau A. 2015. MEG Multivariate Analysis Reveals Early Abstract Action Representations in the Lateral Occipitotemporal Cortex. J

Neurosci. 35:16034–16045.

Vaina LM, Solomon J, Chowdhury S, Sinha P, Belliveau JW. 2001. Functional neuroanatomy of biological motion perception in humans. Proc Natl Acad Sci U S A. 98:11656–11661.

Vangeneugden J, Peelen M V, Tadin D, Battelli L. 2014. Distinct neural mechanisms for body form and body motion discriminations. J Neurosci. 34:574–585.

Vangeneugden J, Pollick F, Vogels R. 2009. Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. Cereb Cortex. 19:593–611.

Wurm MF, Ariani G, Greenlee MW, Lingnau A. 2016. Decoding Concrete and Abstract Action Representations During Explicit and Implicit Conceptual Processing. Cereb Cortex. 26:3390–3401.

Wurm MF, Lingnau A, Wurm XMF, Lingnau A. 2015. Decoding Actions at Different Levels of Abstraction. J Neurosci. 35:7727–7735.

Figure legends

Figure 1 – Action recognition dataset

(a) *We used a dataset of two-second videos depicting five actors performing five actions from five viewpoints. Frames from one example walk video at 90 degrees (top) and one example drink video at 0 degrees (bottom) are shown. We extended this dataset to* ***(b)*** *a “Form only” dataset, containing single (action informative) frames from each two-second movie, and* ***(c)*** *a “Motion only” dataset of point light videos created by labeling joints on actors in each video (a, bottom).*

Figure 2 – Action decoding from video data

(a-b) Within and across view action decoding from MEG data. *We can decode action by training and testing a simple on the same view (‘within-view’ condition), or, to assess viewpoint invariance, training on one view (0 degrees or 90 degrees) and testing on second view (‘across view’ condition), in* ***(a)*** *100 ms overlapping bins (10 ms step size), or* ***(b)*** *10 ms non-overlapping bins. Results are from the average of ten subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding accuracy. Line at bottom of plot indicates group-level significance with p<0.05 permutation test, for the average null distribution across the ten subjects. The first time point in this line is the onset time for each condition, reported in the main text. Inset shows a zoom of decoding time courses from 175-525 ms post-video onset.* ***(c) Action decoding from eye tracking data.*** *We trained a linear classifier on the output of eyetracking data from a separate experiment. We trained the classifier with 80% of the data from all views, and tested on the 20% of held out data. Decoding methods are otherwise analgous to the MEG decoding procedure Results are from the average of five different subjects. Error bars represent standard error across subjects. Horizontal line indicates chance* *decoding (20%). Decoding does not pass the group-level significance threshold of p<0.05 as determined by a permutation test.*

Figure 3 – Confusion matrices for action video dataset. *Confusion matrices for the within and across view decoding conditions in the video dataset for* ***(a)*** *within view decoding at 250 ms post-video onset,* ***(b)*** *across view decoding at 250ms post-video onset,* ***(c)*** *within view decoding at 500ms post-video onset,* ***(d)*** *across view decoding at 500 ms post-video onset,* ***(e)*** *subjects' average behavioral accuracy in Experiment 2. Y-axis shows true action labels and X-axis shows the classifier's prediction (a-d) or subjects' mean response (e). Colorbar indicates the fraction of videos a given action (Y-axis) that was labeled by the classifier or subject as another action (X-axis).*

Figure 4 – Dynamics of action representations. *A temporal cross training matrix showing the decoding results for training a classifier at each point in time (y-axis) and testing the classifier at all other times (x-axis), zoomed in to the time period from 0-1500ms post-video onset, for* ***(a)*** *within-view decoding, and* ***(b)*** *across-view decoding for subjects watching the 2-view video dataset (Experiment 1). Colorbar indicates mean decoding accuracy for ten subjects. Black dots indicate points when decoding is significantly above chance at group level based on p<0.05 significance test. Results along the diagonal for the within and across view decoding are the same as shown in the line plots in Figure 3.* ***(c)*** *Significantly above chance decoding time points, based on a p<0.05 permutation test, for the within view (blue) and across view (red) conditions overlaid on the same plot for the entire time window (-500-3500 ms post video onset).*

Figure 5 – The effects of form and motion on invariant action recognition. (a) *Action can also be decoded invariantly to view from form information alone (static images)* ***(b)*** *Action can*

be decoded from biological motion only (point light walker stimuli). Results are each from the average of ten subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Line at bottom of plot indicates group-level significance with p<0.05 permutation test, for the average null distribution across the ten subjects. The first time point in this line is the onset time for each condition, reported in the main text.

Experiment	Subject	Num. sensors selected (within view)	Num. sensors selected (across view)
video	1	9	11
video	2	7	7
video	3	7	9
video	4	13	18
video	5	6	7
video	6	6	6
video	7	9	10
video	8	7	10
video	9	8	10
video	10	9	12
frame	11	11	11
frame	12	10	4
frame	13	20	27
frame	14	4	5
frame	15	6	6
frame	16	8	9
frame	17	12	19
frame	18	16	23
frame	19	7	7
frame	20	8	10
point light	21	44	62
point light	22	28	20
point light	23	26	36
point light	24	29	32
point light	25	16	24
point light	26	3	3
point light	27	8	11
point light	28	24	25
point light	29	24	21
point light	30	10	15

Table 1 - The average number of sensors selected for decoding (based on a ANOVA on the training data, see Methods) for each of the 10 subjects in each experiment. The entire decoding procedure, including sensor selection is repeated at each time bin. Here we report the average number of sensors selected during the peak decoding time point for each subject.