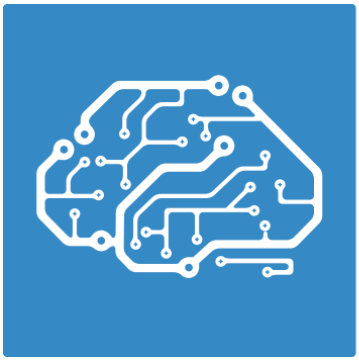


Center for Brains, Minds & Machines

CBMM Memo No. 042 January 6, 2016

Fast, invariant representation for human action in the visual system

by

Leyla Isik*, Andrea Tacchetti*, and Tomaso Poggio

* These authors contributed equally to this work

Abstract:

The ability to recognize the actions of others from visual input is essential to humans' daily lives. The neural computations underlying action recognition, however, are still poorly understood. We use magnetoencephalography (MEG) decoding and a computational model to study action recognition from a novel dataset of well-controlled, naturalistic videos of five actions (run, walk, jump, eat, drink) performed by five actors at five viewpoints. We show for the first time that actor- and view-invariant representations for action arise in the human brain as early as 200 ms. We next extend a class of biologically inspired hierarchical computational models of object recognition to recognize actions from videos and explain the computations underlying our MEG findings. This model achieves 3D viewpoint-invariance by the same biologically inspired computational mechanism it uses to build invariance to position and scale. These results suggest that robustness to complex transformations, such as 3D viewpoint invariance, does not require special neural architectures, and further provide a mechanistic explanation of the computations driving invariant action recognition.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Fast, invariant representation for human action in the visual system

Leyla Isik*, Andrea Tacchetti*, and Tomaso Poggio
Center for Brains, Minds, and Machines, MIT

*These authors contributed equally to this work

Significance Statement

Humans quickly recognize other's actions under many conditions, such as different viewpoints and actors. We investigate how the the recognition of actions is implemented in the brain. We record subjects' brain activity, using magnetoencephalography, a neuroimaging tool, while they watch videos of people performing different actions. Using machine learning we can decode what action the participants viewed based solely on 200 milliseconds of brain activity. This decoding is robust to changes in viewpoint and actor. We used our findings to build a computational action recognition model. We showed that no special circuitry is required to process videos (versus images) or complex transformations like viewpoint (versus simple ones like translation), providing a computational account of action recognition in the brain.

Abstract

The ability to recognize the actions of others from visual input is essential to humans' daily lives. The neural computations underlying action recognition, however, are still poorly understood. We use magnetoencephalography (MEG) decoding and a computational model to study action recognition from a novel dataset of well-controlled, naturalistic videos of five actions (run, walk, jump, eat, drink) performed by five actors at five viewpoints. We show for the first time that actor- and view-invariant representations for action arise in the human brain as early as 200 ms. We next extend a class of biologically inspired hierarchical computational models of object recognition to recognize actions from videos and explain the computations underlying our MEG findings. This model achieves 3D viewpoint-invariance by the same biologically inspired computational mechanism it uses to build invariance to position and scale. These results suggest that robustness to complex transformations, such as 3D viewpoint invariance, does not

require special neural architectures, and further provide a mechanistic explanation of the computations driving invariant action recognition.

Introduction

As a social species, humans rely on the ability to recognize the actions of others in their everyday lives. We can quickly and effortlessly extract action information from rich dynamic stimuli, despite variation in the appearance of these actions due to transformations such as changes in size, viewpoint, actor gait and dynamics (e.g. is this person running or walking towards me, regardless of which direction they are coming from). This ability is paramount to humans' social interactions and even survival. The computations driving this process, however, are poorly understood, as evidence by the fact that humans still drastically outperform state of the art computer vision algorithms on action recognition tasks (1, 2).

Several studies have attempted to define actions and examine which regions in the brain are involved in processing actions and biological motion. In this work we use the taxonomy and definition for actions from (3). Actions are the middle ground between action primitives (e.g. raise the left foot and move it forward) and activities (e.g. playing basketball). Actions are thus possibly cyclical sequences of temporally isolated primitives. In humans and nonhuman primates, the extrastriate body area (EBA) has been implicated in recognizing human form and action (4–6), and the superior temporal sulcus (STS) has been implicated in recognizing biological motion and action (7–9). In humans, the posterior portion of the STS (pSTS) in particular has been found to be involved in recognizing biological motion (10–14). fMRI BOLD responses in this region are selective for particular types of biological motion data in a mirror-symmetric (15) or viewpoint invariant (16) manner. It is behaviorally important to not only recognize actions, but also recognize the actors performing them; recent electrophysiology studies have shown that neurons in macaque STS encode both the identity of an actor invariant to the action they are performing as well as the action being performed invariant to the actor performing it (17). Beyond visual cortex, action representations have been found in human parietal and premotor cortex for performing and viewing certain actions, particularly hand grasping and goal-directed behavior (analogous to monkey "mirror neuron" system) (18). These representations also demonstrate some degree of view tolerance (19), however, recent work suggests that these regions do not code the same abstract concept of action that is found in occipitotemporal regions (20).

Despite progress mapping *where* in the brain actions are represented, only coarse information about the timing of the neural processes, and the underlying computations across the brain is available. Here we will, look at responses to natural movies, we will investigate the dynamics of neural processing to help elucidate the underlying neural computations, and finally, we will implement these insights into a biologically-inspired computational model that performs action recognition from naturalistic videos. Specifically, we use magnetoencephalography (MEG) decoding analysis and a computational model of the visual cortex, to understand when and how different computations are carried out to perform actor and view invariant action recognition in the visual system.

We showed with MEG decoding that the brain computes a representation for actions very quickly (in under 200 ms after the video onset) and that this early representation is invariant to non-affine transformations (view and actor). We next used these insights to extend a computational and theoretical framework for invariant object recognition in still images to recognize actions from videos in a manner that is also invariant to actor and viewpoint on the same dataset. Finally we also show, using behavioral data, MEG, and the model, that both form and motion are crucial for action recognition and that different computational processes are recruited to make sense of form-depleted or motion-depleted stimuli.

Results

Novel invariant action recognition dataset

To study the effect of changes in view and actor on action recognition, we filmed a dataset of five actors performing five different actions (drink, eat, jump, run and walk) on a treadmill from five different views (0, 45, 90, 135, and 180 degrees from the front of the actor/treadmill; the treadmill rather than the camera was rotated in place to acquire from different viewpoints) [Figure 1]. The dataset was filmed on a fixed, constant background. To avoid low-level object/action confounds (e.g. the action “drink” being classified as the only videos with water bottle in the scene) the actors held the same objects (an apple and a water bottle) in each video, regardless of the action they performed. This ensures that the main variations between videos are the action, actor, and view, and allows controlled testing of different hypotheses concerning when and how invariant recognition arises in the human visual system. Each action-actor-view combination was filmed for at least 52-seconds. The videos were 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 mid air or on the ground). The dataset includes 26 two-second clips for each actor, action, and view, for a total of 3250 video clips. This dataset allows testing of actor and view invariant action recognition, with few low-level confounds. A motion energy model (C1 layer of the model described below) cannot distinguish action invariant to view [Supplemental Figure 1].

Figure 1

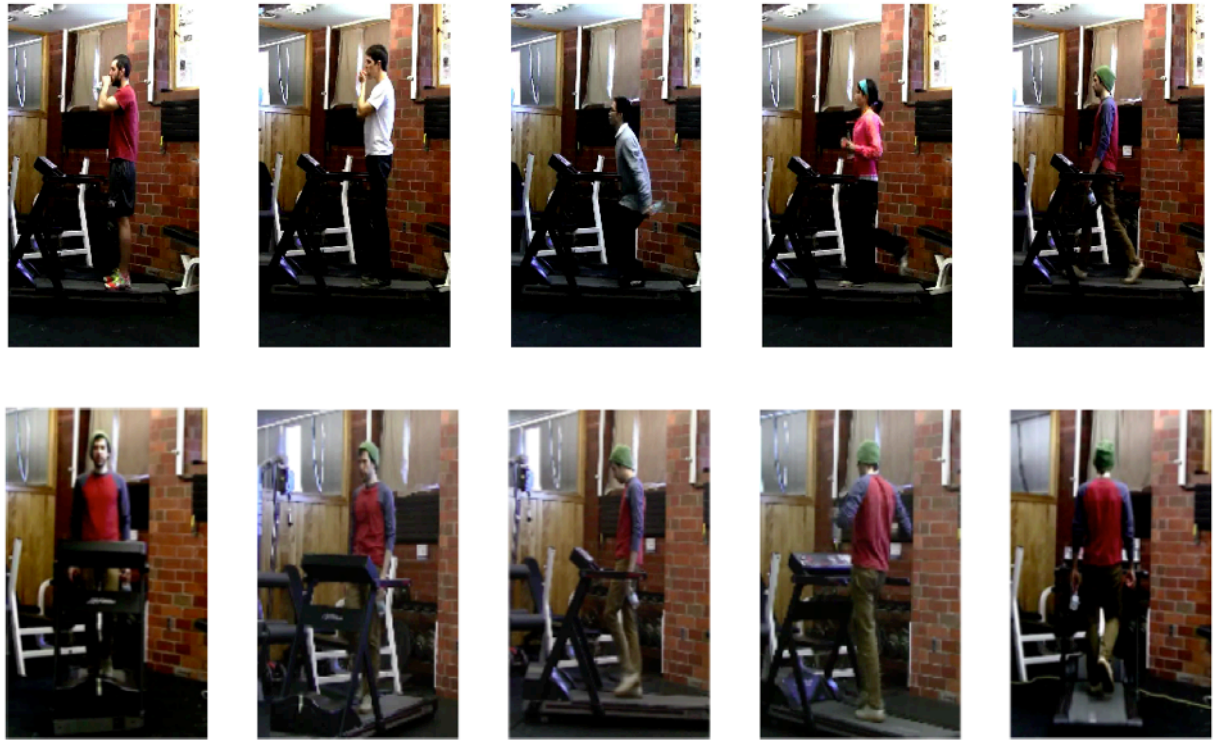


Figure 1: Novel action recognition dataset. *The dataset consists of five actors, performing five actions (drink, eat, jump, run and walk), at a fixed position in the visual field (while on a treadmill) and on a fixed background across five different views (0, 45, 90, 135, and 180 degrees). To avoid low-level confounds, the actors held the same objects in each hand (a water bottle and an apple), regardless of the action performed.*

Readout of actions from MEG data is early and invariant

Eight subjects viewed two views (0 and 90 degrees) from the above dataset and were instructed to recognize which of the five actions was performed in each video clip while their neural activity was recorded in a MEG scanner. We use decoding analysis, which applies a linear machine learning classifier to discriminate stimuli based on the neural response they elicit, to analyze the MEG signals. By repeating the decoding procedure at each 10 ms time step, we can see when

different types of stimulus information are present in the brain. Action can be read out from the subjects' MEG data as early as 200 ms after the video starts (after only about 6 frames of each two-second video) [Figure 2]. This is surprising, given that 200 ms was much less than a cycle of most actions, suggesting that the brain can compute a representation for these actions from different partial sequences of each.

We can test if these MEG signals are invariant to view by training the classifier on data from subjects viewing actions performed at one view and testing the classifier on a second held out view. We decoded by training only on one view (0 degrees or 90 degrees), and testing on a second view (0 degrees or 90 degrees). There is no difference in the latency between the 'within view' case (train and test at 0, or train and test at 90) and the 'across view' case (train on 0 and test 90, or train on 90 and test on 0) [Figure 2], suggesting that the MEG signals can generalize across views within 200 ms. Actor-invariant signals can be similarly read out at 200ms [Supplemental Figure 2], and subjects' eye movements cannot account for this decoding [Supplemental Figure 3]. These experiment show that the human visual system computes a representation for actions that we are able to read out from their MEG signals. This representation is computed very quickly (200ms) and is immediately invariant to changes in actor and 3D rotation.

Figure 2

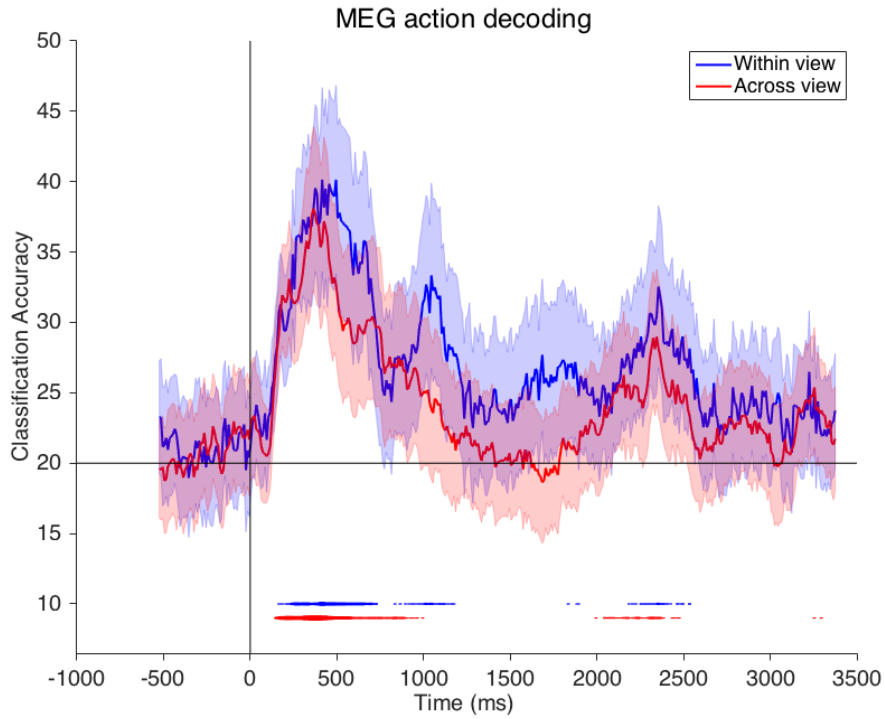


Figure 2: Decoding action from MEG data. *We can decode action by training and testing on the same view ('within-view' condition), or, to test viewpoint invariance, training on one view (0 degrees or 90 degrees) and testing on second view ('across view' condition). Results are each from the average of eight different subjects. Error bars represent standard deviation. Horizontal line indicates chance decoding accuracy (see Supplementary Materials). Lines at the bottom of plot indicate significance with p<0.01 permutation test, with the thickness of the line indicating if the significance holds for 2-8 subjects.*

Recognizing actions with a biologically-inspired hierarchical model

In order to provide a mechanistic explanation of how the brain quickly computes a representation for action that is invariant to viewpoint and actor we implemented a computational model that recognizes actions from videos. This model is a an extension of a class of computational models of visual cortex, convolutional neural networks, which have successfully explained object recognition from static images (21–23), to stimuli that extend in time. The model's structure is hierarchical: the input video goes through a layer of computation and the output of this layer serves as input to the next layer, the sequence of layers is inspired by Hubel and Wiesel's findings in primary visual cortex, and is constructed by alternating layers

of simple cells, which perform a template matching or convolution, and complex cells, which perform max pooling (24). The specific model that we present here consists of two simple-complex layer pairs [Figure 3a]. Further, our model directly implements insights from the MEG timing data: it is completely feed-forward, to account for the fast MEG readout, and further it generalizes across 3-D viewpoint transformations at the same hierarchical layer and using the same computational mechanism it employs to generalize across changes in position and scale, to account for the fact that early MEG signals were invariant to 3-D viewpoint.

Qualitatively, the model works by detecting the presence (or lack thereof) of a certain video segment (a template) in the input stimulus. The exact position in space and time of the detected template is discarded by the pooling mechanism and only the information about its presence is passed on to the next layer. Our model shares a basic architecture with deep convolutional neural networks. Notably, it is designed to closely mimic the biology of the visual system. It has few layers and hard coded S1 templates (moving Gabor-like stimuli, with both a spatial and temporal component, that model the receptive fields found in primate V1 and MT (25–27)). Our model offers an interpretable mechanism that explains the underlying computations [Figure 3a].

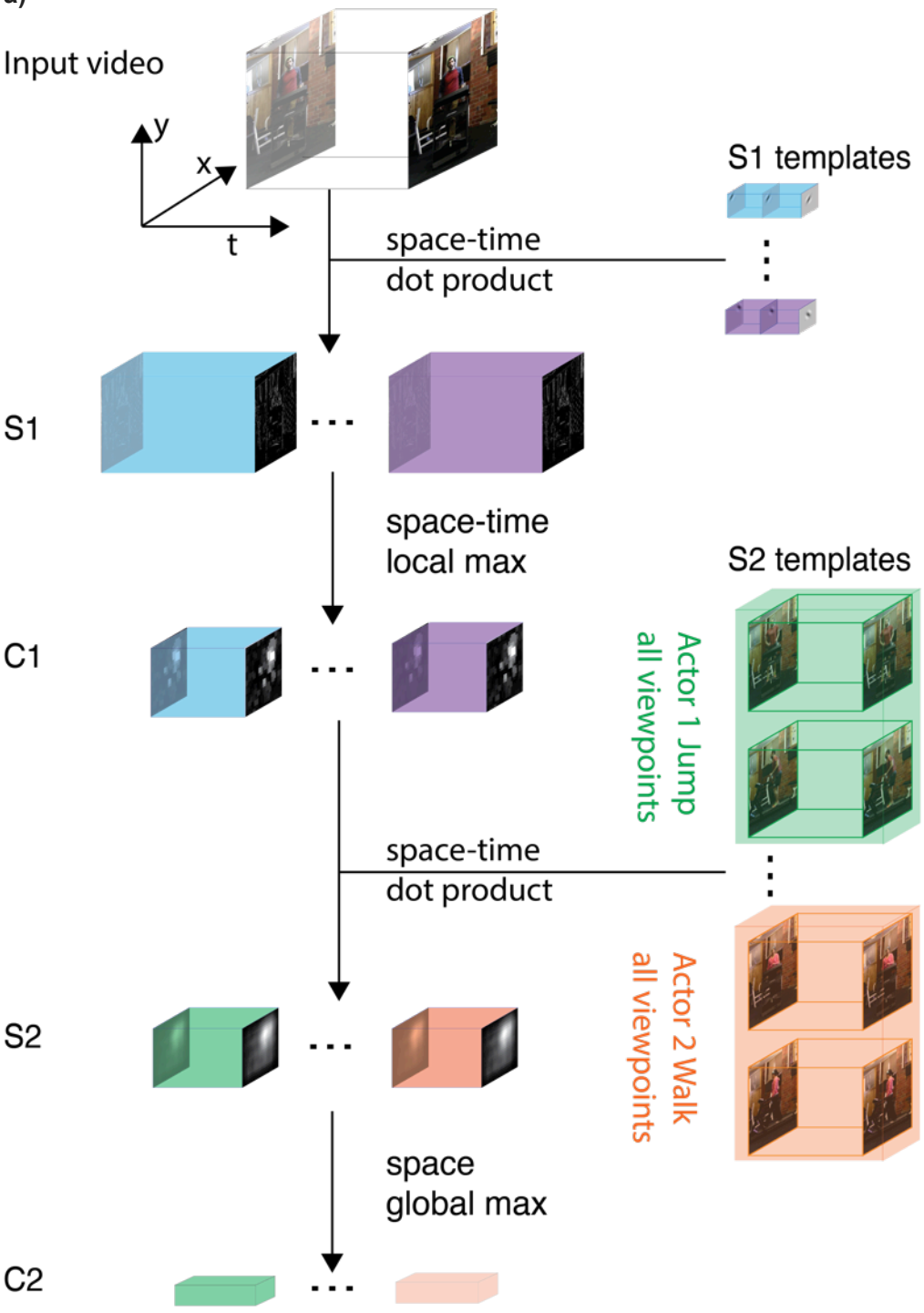
In order to produce a response that is invariant to rotation in depth, the model's top complex cell units (C2) pool over all templates containing patches of videos of a single actor performing a specific action recorded at different viewpoints. This novel "pooling across channels" mechanism detects the presence of a certain template (e.g. the torso of someone running) regardless of its 3D pose. Many theories and experimental evidence have suggested how this wiring across views is learned in development (28–31). We compare this structured model to an unstructured control model, which contains the same templates, but where action is not taken into account in the pooling scheme and instead each C2 cell pools over a random, unstructured set of S2 cell templates [Figure 3b].

We test the performance of our model by training and testing a machine learning classifier to recognize actions based on the model output. We show that the simple pooling mechanism just described is sufficient to account for viewpoint invariance. Both the model with structured connectivity pattern and the model with a unstructured connectivity can recognize action when training and testing of the machine learning classifier happens within one viewpoint (82+/-7% and 79+/-5% accuracy +/- standard deviation, respectively). However, the model with structured pooling provides significantly better accuracy on the view-invariant action recognition task (49

+/-5% vs. 36+/-5% accuracy) [Figure 4] when the machine learning classifier is trained on videos at one of two viewpoints, 0 or 90 degrees and tested at the opposite one. In addition, the classifier is always tested on model responses to videos from a held-out actor, so, like the MEG data, the model can also recognize actions invariant to actor.

Figure 3

a)


Input video
y
x
t
S1 templates
space-time
dot product
S1
space-time
local max
S2 templates
C1
Actor 1 Jump
all viewpoints
space-time
dot product
Actor 2 Walk
all viewpoints
S2
space
global max
C2

b)

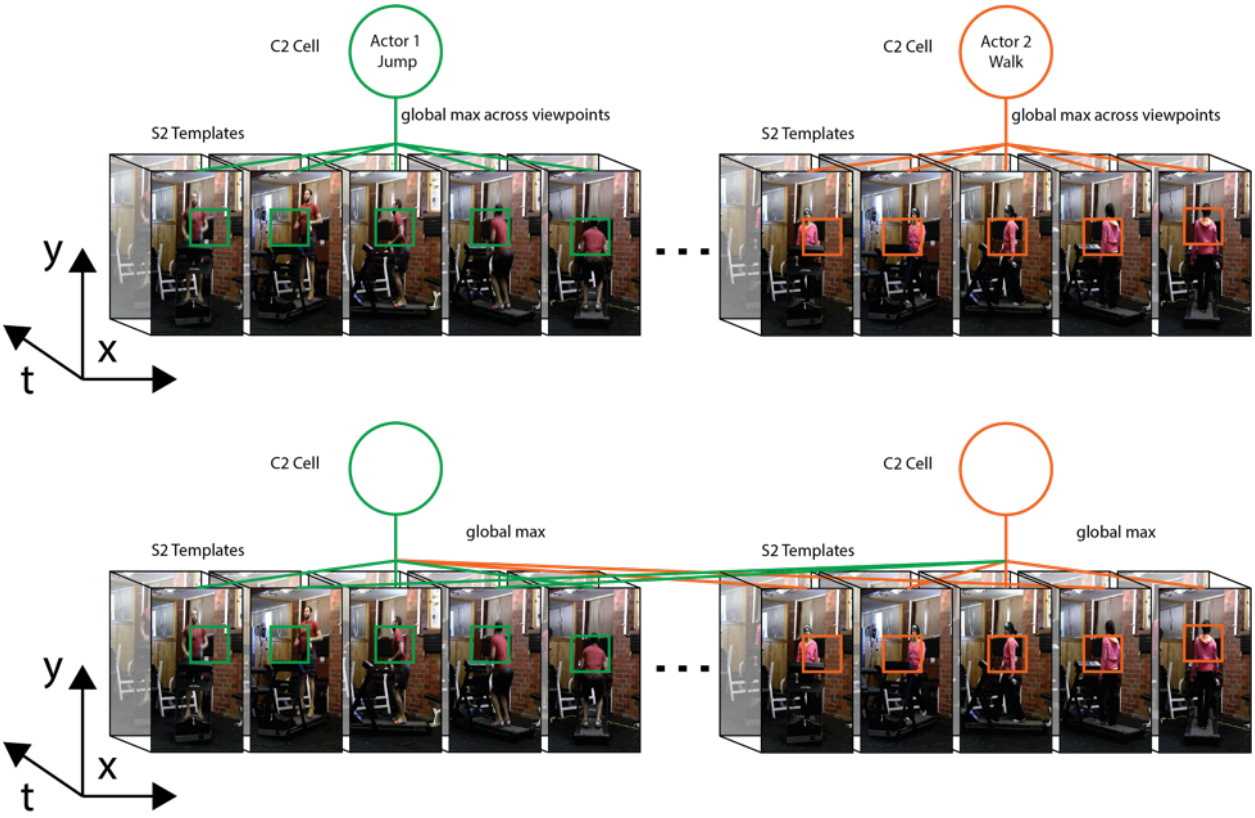


Figure 3: Biologically inspired computational model for action recognition. ***(a)*** *An input video is convolved with moving Gabor like templates at the S1 layer. At the C1 layer, a local max pooling is applied across position and time. At the S2 layer, previous layer's outputs are convolved with templates sampled from a sample set of videos disjoint from the test set. Videos in the sample set go through the S1 and C1 layers before being used for sampling. At the final layer a global max across positions and views is computed.*

(b) *Invariance is obtained by enforcing structure in the wiring between simple and complex cells at the S2-C2 pooling stage. C2 units pool over all S2 units whose templates come from videos containing a particular actor performing a particular action across different views. We compare this experimental model [top] to an unstructured control model [bottom], which contains the same S2 templates, but where each C2 cell pools over a random, unstructured set of S2 cell templates.*

Figure 4

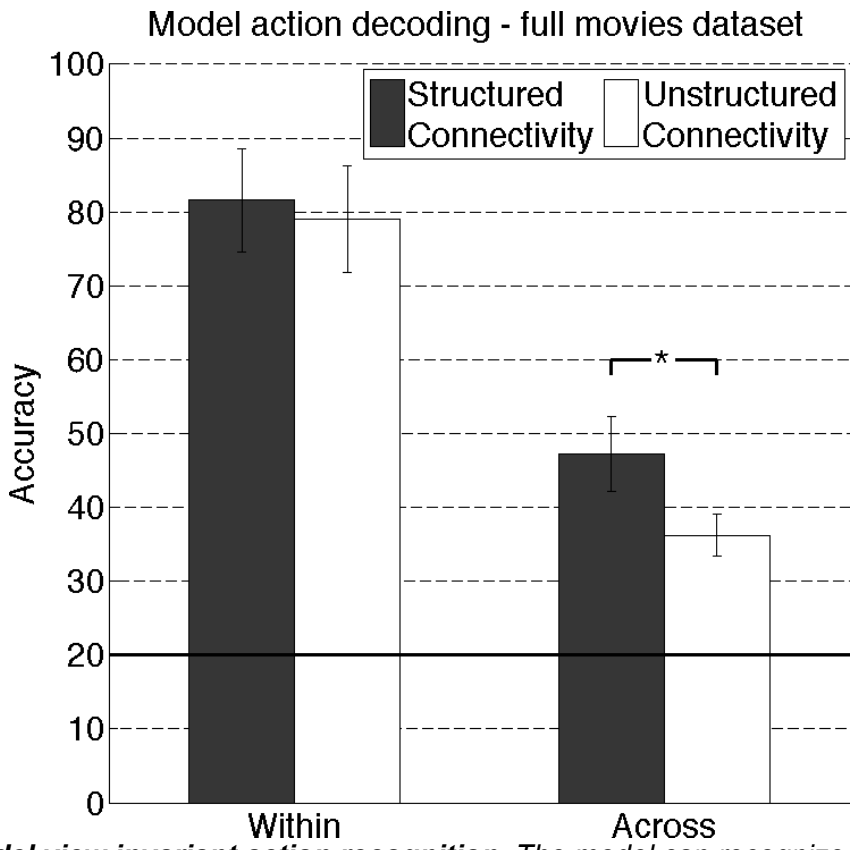


Figure 4: Model view invariant action recognition. *The model can recognize action when trained and tested on the same view ('within-view' condition), or trained on one view (0 degrees or 90 degrees) and tested on second view ('across view' condition). The Experimental model employs structured pooling as described in Figure 3b, top, and the Control model employs random C2 pooling as described in Figure 3b, bottom. Error bars indicated standard deviation across model runs [see supplementary information]. Horizontal line indicates chance performance (20%). Asterisk indicates a statistically significant difference with p<0.01.*

The roles of form and motion in invariant action recognition

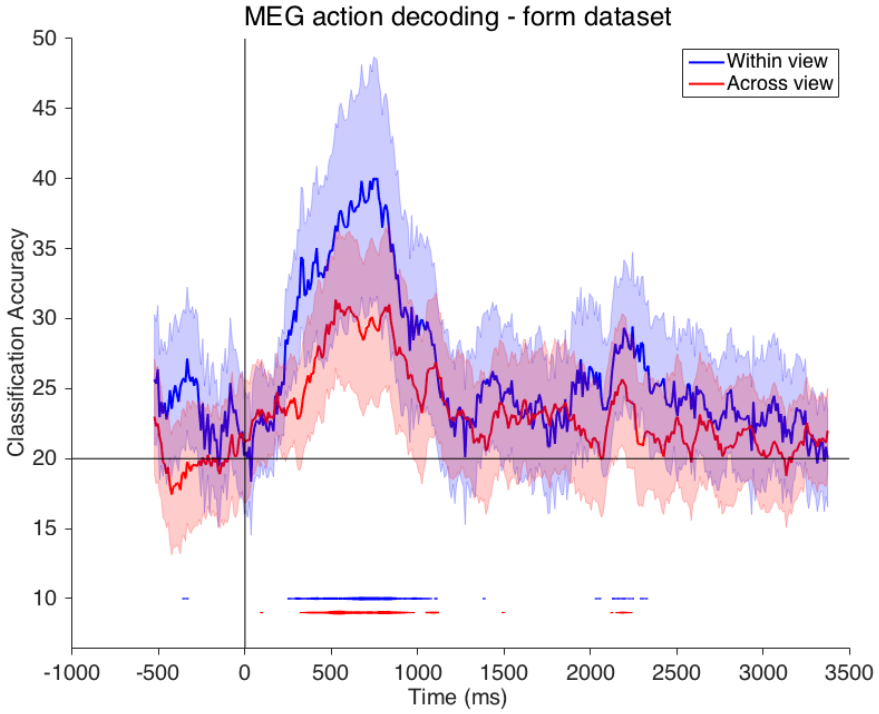
To test the effect of form and motion on action recognition, we used two limited stimulus sets. The first 'Form' stimulus set consisted of one static frame from each video (no motion information). The second 'Motion' stimulus set, consisted of point light figures that are comprised of dots on each actor's head, arm joints, torso, and leg joints and move with the actor's joints (limited form information) (32).

Eight different subjects viewed each of the form and motion datasets in the MEG. We could decode action within view in both datasets. Decoding performance across view, however, was significantly lower than the case of full movies [Figure 6a-b]. In addition, subjects' behavioral performance dropped from 92 +/-4% correct with full movies to 76 +/-11% correct on the 'Form' dataset and 78 +/-18% on the 'Motion' dataset, suggesting that the lack of motion information hinders recognition and this recognition deficit is reflected particularly in the MEG results.

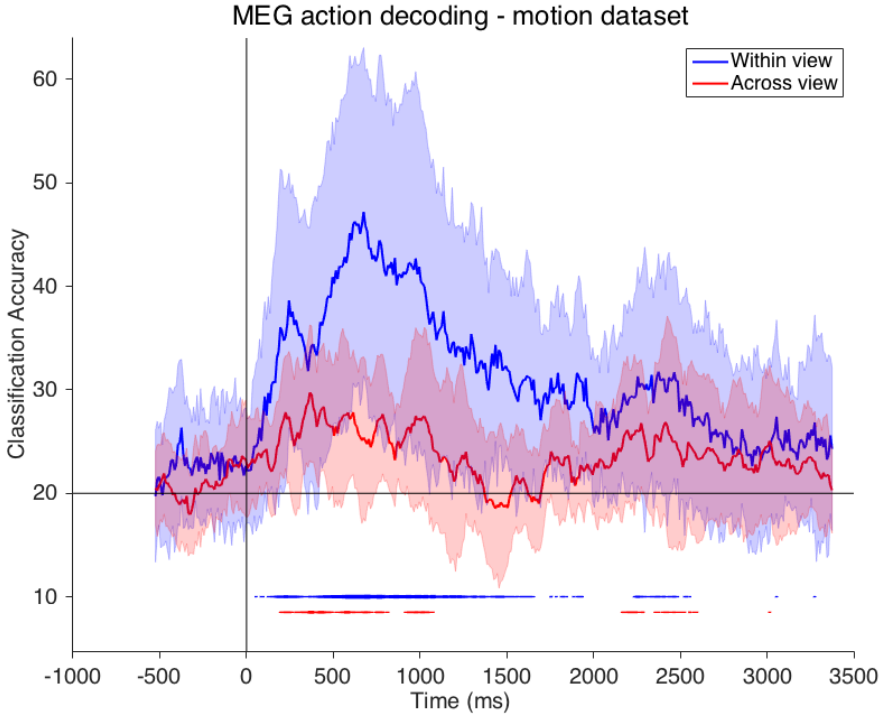
We examined the effects of form and motion with our model by testing both stimulus sets on a model with templates sampled from full videos. While it is still possible to classify correctly which action was performed, performance was significantly lower than in the case where full videos were used. The experimental model (with S2 to C2 pooling over templates that are rotated in depth) outperforms the control model (where the wiring is randomized in classifying actions from static frames) [Figure 6c]. Both the experimental model with a structured pooling pattern and the control model are completely unable to generalize across viewpoint on form-depleted stimuli. Both models, however, are able to generalize across actors if trained and tested on the same view [Figure 6d]. The MEG and model results suggest that form and motion are both critical in recognizing actions, particularly in a view invariant manner.

Figure 5

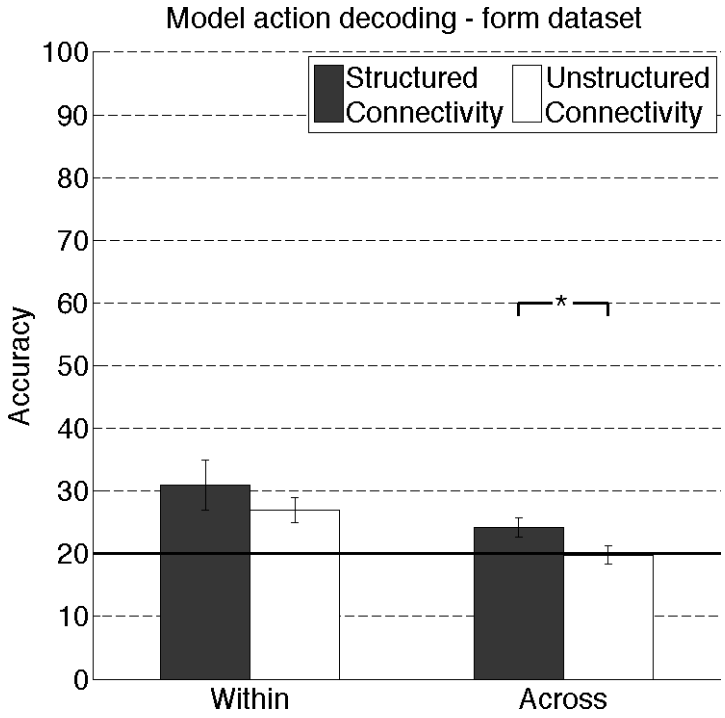
a)


MEG action decoding - form dataset
Within view
Across view
Classification Accuracy
50
45
40
35
30
25
20
15
10
-1000
-500
0
500
1000
1500
2000
2500
3000
3500
Time (ms)

b)


MEG action decoding - motion dataset
Within view
Across view
Classification Accuracy
60
50
40
30
20
10
-1000
-500
0
500
1000
1500
2000
2500
3000
3500
Time (ms)

c)


Model action decoding - form dataset
Structured Connectivity
Unstructured Connectivity
Accuracy
100
90
80
70
60
50
40
30
20
10
0
*
Within
Across

d)

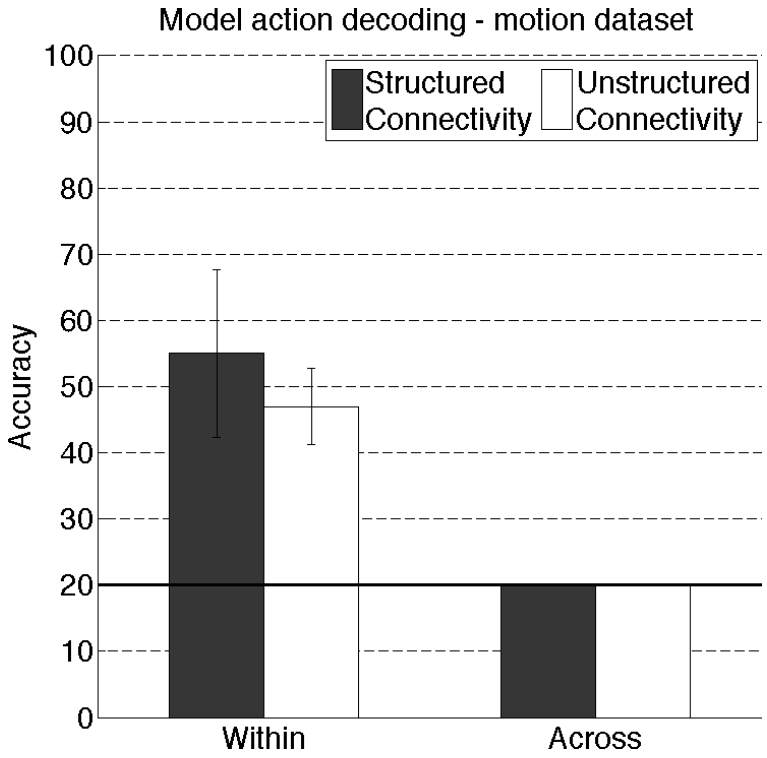

Model action decoding - motion dataset
Structured Connectivity
Unstructured Connectivity
Accuracy
100
90
80
70
60
50
40
30
20
10
0
Within
Across

Figure 5: The effects of form and motion on invariant action recognition (a) *Action can also be decoded invariantly to view from static images.* ***(b)*** *Action can be decoded from biological motion only (point light walker stimuli). Results are each from the average of eight different subjects. Error bars represent standard deviation. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with p<0.01 permutation test, with the thickness of the line indicating if the significance holds for 2-8 subjects.* ***(c)*** *The model can recognize action from static frames, but the performance is much lower than with full videos. The model with structured connectivity employs the pooling pattern across views shown in [Figure 3b, top], and the model unstructured connectivity employs random C2 pooling as described in [Figure 3b, bottom]. Error bars indicated standard deviation across model runs [see supplementary information]. Horizontal line indicates chance performance (20%). Asterisk indicates a statistically significant difference with p<0.01.* ***(d)*** *The model can recognize actions from form-depleted stimuli. The classification accuracy is significantly lower than what can be obtained with full videos. The model is unable to generalize across views with motion information.*

Discussion

We analyzed the dynamics of invariant action recognition in the human brain to find that action recognition occurs as early as 200 ms after a video begins. This early neural representation is invariant to changes in actor and 3-D viewpoint. These timing results provide compelling evidence that the bulk of these computations are performed in feedforward manner, and interestingly that invariant representations for action are computed at same time as non-invariant representations. This seems to be in contrast to object recognition where invariance increases at subsequent layers in the ventral stream (33–35) causing a delay in decoding accuracy (36).

Action recognition occurs on a similarly fast time scale to, but slightly later than, the previously documented timing of object recognition in the human brain (60 ms for non-invariant object recognition and 100-150 ms for size and position object recognition) (36–40). The slightly later timing of action recognition relative to object recognition and the fact that invariant and non-invariant action representations have the same latency suggest higher level visual features (requiring the integration of both form and motion cues) are recruited for even basic action recognition, in contrast to simple object recognition (which is based on low-level features like lines and edges). This is consistent with our finding that early action representations (unlike previously reported early object signals (36)) are already invariant, as well as previous MEG decoding object recognition results showing that more abstract categorizations take more time (38, 39).

We used these neural insights to develop a feedforward cortical model that recognizes action invariant to actor and view (non-affine transformations). Inspired by the MEG timing data, the computations underlying the model's invariance to complex transformations are performed in the same model layer and using the same pooling mechanism as size and position (affine transformations). Our modeling results offer a computational explanation of the underlying neural mechanisms that lead to the fast and invariant action representations in visual cortex. In particular, our model showed that a simple-complex cell architecture (41), is sufficient to explain fast invariant action recognition across video stimuli with complex transformations, suggesting that no special neural circuitry is required to deal with non-affine transformations. The model architecture is inspired by prior work in modeling the recognition of biological motion (42) and unlike previous extensions of object recognition systems to actions in videos (43) is able to generalize across 3D rotation and actors in realistic, complex videos. The mechanism employed

to achieve invariance to non-affine transformations by pooling over them has been proposed for face 3D rotation in artificial stimuli (44), the work presented here extends that framework for the first time to natural stimuli that extend in both space and time.

The highest performing computer vision systems on action recognition tasks are deep convolutional neural networks, which have a similar architecture to our model, but more layers and free parameters that are tuned for performance on a given classification task using backpropagation (1). Our model is designed to have biologically faithful parameters and mimic human visual development and prioritizes interpretability of its underlying computational mechanisms. This modeling effort is primarily concerned with understanding and modeling how the brain accomplishes action recognition, rather than creating an artificial system that maximizes performance through a non-biological method.

We found that biological motion and form are each enough alone to recognize actions, however decoding and model performance for the viewpoint invariant task drops to almost chance when either form or motion information is removed. This is also reflected in a slight drop in behavioral performance. While form- or motion-depleted data sets afford more experiment control and have been the focus on much prior study, it is worth considering if they are the best way to understand the neural mechanisms underlying action recognition. Humans can indeed recognize action from diminished stimuli, but here we show it elicits different neural response than full video stimuli, particularly in the case of viewpoint invariant recognition. Moving toward more naturalistic stimuli, possibly in conjunction with controlled experiments with form or motion-only data, is important to understand the full range of neural responses that support human action recognition.

Conclusions

This work shows that neural representations for actor and view invariant action recognition are computed remarkably quickly, within 200ms. These timing insights were directly used to influence the structure of a computational model, namely its feedforward architecture and pooling across viewpoint in same layer as affine transformations, to perform view invariant action recognition. This model provides a computational explanation of our MEG timing results, as well as an interpretable alternative to deep convolutional neural networks. Close interchange between artificial intelligence and neuroscience efforts will help move towards a deeper

understanding of realistic perception of humans' actions and advance the design artificial systems that understand our actions and intentions.

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 gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for 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.

Online Methods

Action recognition dataset

We filmed a dataset of five actors performing five actions (run, walk, jump, eat and drink) from five views (0, 45, 90, 135, and 180 degrees from the front) on a treadmill in front of a fixed background. By using a treadmill we avoided having actors move in and out of frame during the video. To avoid low-level object confounds, the actors held a water bottle and an apple in each hand, regardless of the action they performed. Each action was filmed for 52 seconds, and then cut into 26 two-second clips at 30 fps.

For single frame dataset, single frames that were as unambiguous as possible for action identity were hand selected (special attention was paid to actions eat and drink and occluded views). For the motion point light dataset, the videos were put on Amazon Mechanical Turk and workers were asked to label 15 points in every single frame: center of head, shoulders, elbows, hands, torso, hips, knees, and ankles. The spatial median of three independent labeling of each frame was used to increase the signal to noise ratio. The time series for each of the 15 points was independently low-passed to reduce the high frequency artifacts introduced by the single-frame labeling we used

MEG experimental procedure

Twenty subjects age 18 or older with normal or corrected to normal vision took part in the experiment. The MIT Committee on the Use of Humans as Experimental Subjects approved the experimental protocol. Subjects provided informed written consent before the experiment. One subject (S5) was an author, and all others were unfamiliar with the experiment and its aims.

In the first experiment, eight 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. In the second experiment, eight subjects were shown 50 static images, which were single frames from the videos in Experiment 2, for 2 seconds presented 20 times each. In the third experiment, eight subjects were shown 10 two-second video clips, which consisted of point-light walkers traced along one actor's videos in experiment two, 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 (twice for each of the fifty videos or images in each experiment) 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 each trial to avoid systematic motor confounds in the decoding.

The videos were presented using Psychtoolbox to ensure accurate timing of stimulus onset. Each video had a duration of 2s and had a 2s ISI. 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 was collected using an Elekta Neuromag Triux scanner with 102 magnetometers at 204 planar gradiometers. The MEG data were sampled at 1,000 Hz. The signals were pre-processed using and preprocessed using Brainstorm software(45). First the signals were filtered using temporal Signal Space Separation (tSSS) with Elekta Neuromag software. Next, Signal Space Projection (SSP) (46) was applied for movement and sensor contamination, and the signals were band-pass filtered from 0.1–100 Hz to remove external and irrelevant biological noise (47, 48). Finally the MEG data was divided into epochs from -500–3500 ms, relative to video onset. SSP, bandpass filtering and epoching were applied using Brainstorm software.

Eyetracking

To verify that the subjects' eye movement could not account for the action discrimination, eye tracking was performed during MEG recordings while five subjects viewed the entire 125 image dataset Supplemental Experiment 1 (subjects S1-S5 viewing five actors performing five actions at five views) with the Eyelink 1000 eye tracker from SR Research. A nine-point calibration was used at the beginning of each experiment. We then performed decoding using the position data for the left and right eye, and found that decoding performance was not significantly above chance for more than two consecutive 5ms time bins, much below the significance threshold outlined for decoding (Supplemental Figure 3).

MEG decoding analysis methods

MEG decoding analyses were performed with the Neural Decoding Toolbox (49), 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 (or actor) 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.

The time series data of the magnetic field measure in each sensor (for both 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. Decoding analysis was performed using cross validation, where the classifier was trained on a randomly selected subset of 80% of data for each stimulus and tested on the held out 20%, to assess the classifier's decoding accuracy.

To improve signal to noise, we averaged the different trials for each stimulus in a given cross validation split. We next Z-score normalized that data and performed sensor selection using the training data. We performed sensor selection by applying an ANOVA to each sensor separately using data from the training set only, to choose sensors selective for stimulus identity with $p<0.05$ significance based on F-test. Decoding analyses were performed using a maximum correlation coefficient classifier, which computes 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 is assigned the label of the class of the training data with which it is maximally correlated.

We repeated the above decoding procedure over 50 cross validation splits, at each time bin to assess the decoding accuracy versus time. Decoding accuracy is reported as the average percent correct of the test set data across all cross validation splits.

We assessed decoding significance using a permutation test. To perform this test, we generated a null distribution by the performing the above decoding procedure for 100 time bins using data with randomly shuffled labels. Specifically, the five action labels within each viewpoint were shuffled, and thus exchangeable under the null hypothesis, and the shuffling was performed once and fixed for all cross-validation runs at each time bin. We recorded the peak decoding accuracy for each time bin, and used the null distribution of peak accuracies to select a threshold where decoding results performing above all points in the null distribution for the corresponding time point were deemed significant with $P < 0.01$ (1/100). The first time decoding reached significantly above chance ("significant time") was defined as the point when accuracy was significant for five consecutive time bins. This significance criterion was selected such that no spurious correlations in the baseline period were deemed significant. We compared the onset latencies of the within and across view time courses for decoding actions from full videos by examining the single subject onset latency difference (within-view minus across-view latency) and modeling subject as a fixed effect.

See Isik *et al.* 2014 for more decoding methods details (36).

Model

The model was written using the CNS: Cortical Network Simulator (23) and is composed of 4 layers. The input video is scaled down, preserving the aspect ratio, with the largest spatial dimension being 128px. A total of three scaled replicas of each video are run through the model in parallel; the scaling is by a factor of 1/2. The first layer is composed of a grid of simple cells placed 1px apart (no sub-sampling), the templates for these units are Gabor receptive fields that move in space while they change phase (as described in previous studies on the receptive fields of V1 and MT cells (25, 27)). Cells have spatially square receptive field of size 7, 9 and 11px, extend for 3, 4 and 5 frames and compute the dot product between the input and their template. The Gabor filters in each receptive field move exclusively in the direction orthogonal to the spatial modulation at 3 speeds, linearly distributed between 4/3 and 4 pixels per frame. The second layer (C1) is a grid of complex cells that compute the maximum of their afferent simple cells. Cells are placed 2 units apart in both spatial dimensions (spatial subsampling by a factor

of 2) and every unit in the time dimension (no time subsampling). Complex cells at the C1 level have spatial receptive fields of 4 simple cells and span 2 scales with one scale overlap, bringing the number of scaled replicas in the model from 3 to 2. The third layer (S2) is composed of a grid of simple cells that compute the dot product between their input and a stored template. The templates at this level are sampled randomly from a sample dataset that has no overlap with the test set; we sample 512 different templates, uniformly distributed across classes and across videos within each class. The cells span 9, 17 and 25 units in space and 3, 7 and 11 units in time. The fourth layer, C2, is composed of complex units that compute the maximum of their inputs; C2 cells pool across all positions and scales. The wiring between simple and complex cells at the C2 layer is described by a matrix with each column corresponding to a complex cell; each column is then a list of indices for the simple cells that the complex cells pools over. in the structured models each column indexes cells with templates samples from videos featuring a single actor performing a single action. In control models, the rows of this matrix are scrambled and therefore the columns (i.e. indices of simple cells pooled together by a single complex cell) have no semantic common thread. S2 template sizes are always pooled independently from one another. The output of the C2 layers is concatenated over time and cells and serves as input to a supervised machine learning classifier.

Video pre-processing and classification

We used non-causal temporal median filtering background subtraction for all videos (50). All classification experiments for the model were carried out using the Gaussian Kernel Regularized Least Squares classification pipeline available in the GURLS package (51)**.** Both the kernel bandwidth and the regularization parameter were chosen using leave-one-out cross validation.

Model experiments

Model experiments are divided in three steps: sampling templates from a sample set in order to populate the model's S2 units, computing the model's response to a set of training and test videos and lastly training and testing a classifier on these responses to report its accuracy. For each of the experiments reported the computer vision model was an instance of the general architecture outlined above and the sample, training and test set were a subset of the dataset described in the main text. A few details were modified for each task in the S2 and C2 layers to make sure the model tested the hypothesis we set forward in that particular experiment and to

avoid having S2 templates sampled from the test set. For the same reasons, we used different set of videos for each experiment. Here we describe these slight modifications.

For the action recognition task [Figure 4], templates were sampled from a sample set containing videos of four out of the five actors performing all five actions and at all five views. In the experimental model, all S2 cells of the same size, with templates sampled from videos of the same actor-action pair (regardless of viewpoint) were wired to the same C2 cell yielding a C2 layer composed of 60 complex cells. In the control model we scrambled the association between templates and videos of origin (after sampling). The training set for this experiment was composed of 600 videos of four of the five actors performing all five actions at either the frontal or side viewpoints. The test set was composed of 150 videos of the fifth actor performing all five actions at either the frontal or side viewpoint. We only used either one of the viewpoints to train or test so as to verify the ability of the model to recognize actions within the same view and to generalize across views. This sample/train/test split was repeated five times, using each actor for testing once and re-sampling the S2 templates each time.

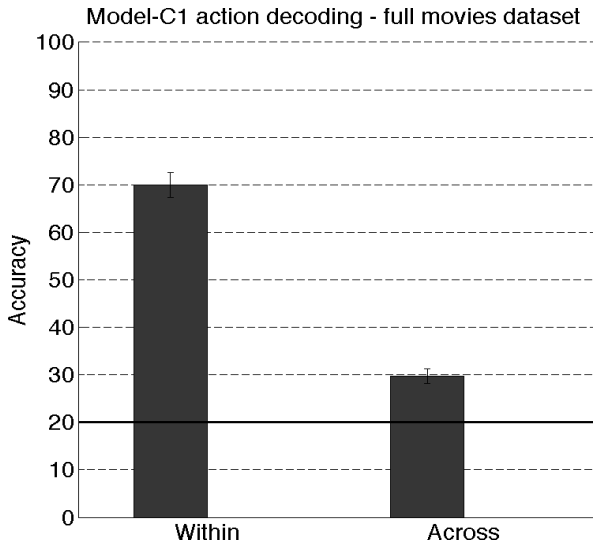
For the actor recognition experiment [Figure 5b], templates were sampled from a sample set containing videos of three of the five actors performing all five actions at all five views. In the experimental model, all S2 cells of the same size, with templates from the videos of the same actor-viewpoint pair (regardless of action), were wired to the same C2 cell yielding a C2 layer composed of 45 complex cells. The training set for this experiment was composed of 600 videos of the two held out actors performing four of the five actions at all viewpoints. The test set was composed of 150 videos of the two left out actors performing the fifth action at all five viewpoints. The experiment was repeated five times changing the two actors that were left out for identification and the action used for testing, the S2 templates were re-sampled each time.

The form only classification experiment [Figure 6c] was conducted using the method described above for the action recognition experiment with the only difference that the test set was composed of videos that only featured one frame repeated for the entire duration of the clip. The motion only classification experiment was also conducted using the method described above for the action recognition experiment with the only differences being that only 100 form depleted videos of the held out actor were used for testing and that only 40 from depleted videos were used for training. Furthermore the experiment was not repeated using different actors for the test phase due to the prohibitive cost of acquiring human annotation for joint location in each

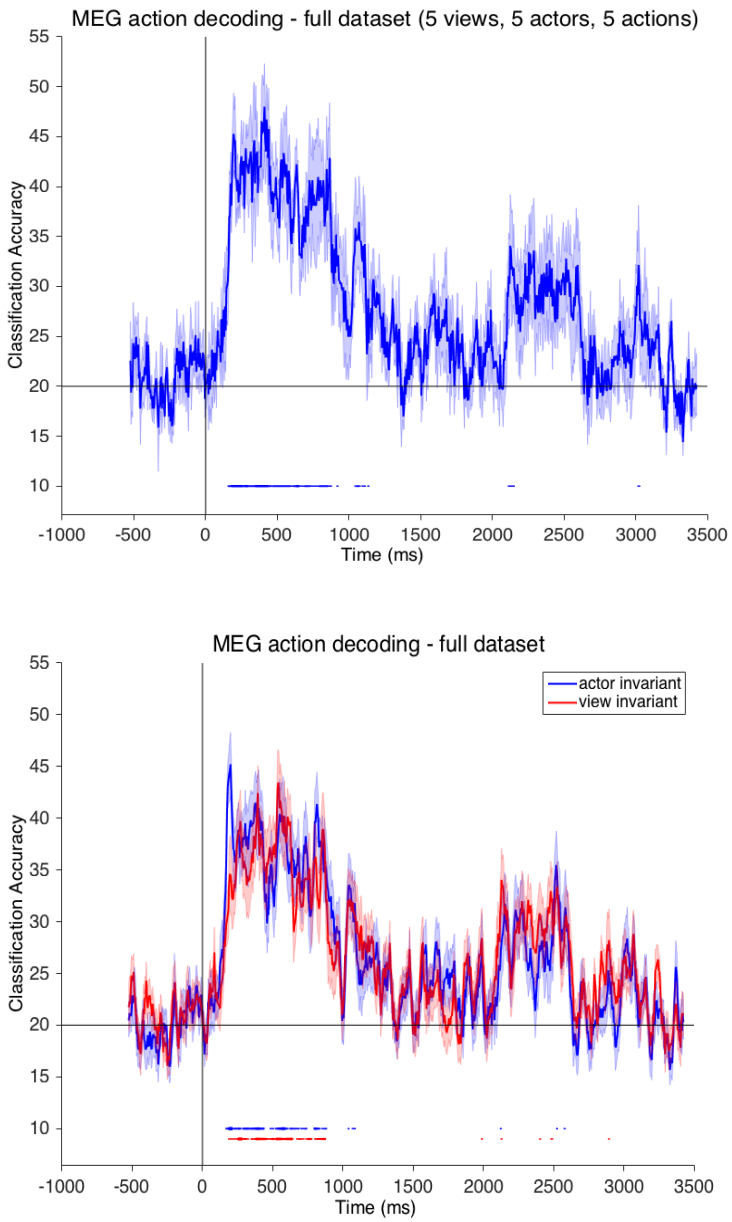
frame (see Dataset methods above), however the experiment was repeated using three distinct instances of our model for each of which the S2 templates were independently sampled.

Supplemental figures

Supplemental Figure 1

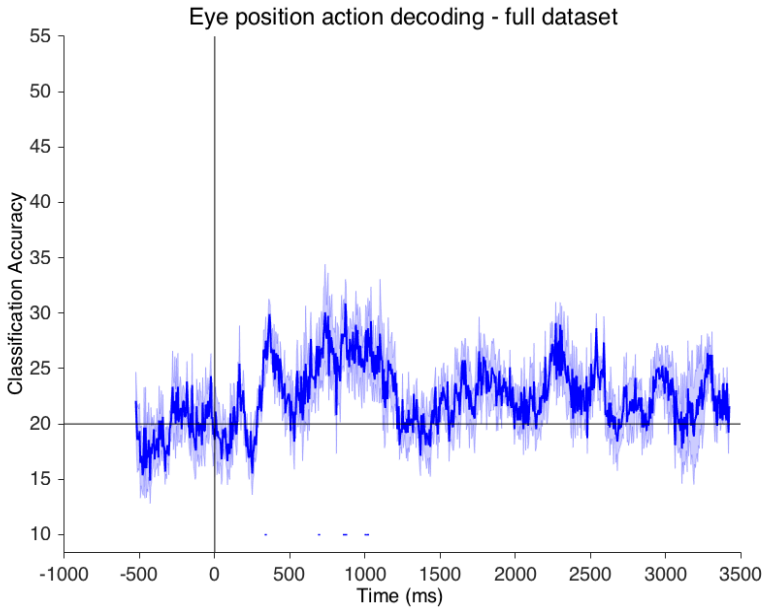


Supplemental Figure 1: C1 performance on dataset. *The output of the C1 layer of the model (analogous to V1-like model) cannot classify action invariant to viewpoint ('Across' condition). Error bars indicated standard deviation across model runs. Horizontal line indicates chance performance (20%).*



Supplemental Figure 2 - MEG decoding from all five views and five actors in dataset. ***(a)*** *Action can be decoded from subject's MEG data as early as 200 ms after stimulus onset (time 0).* ***(b)*** *Action can be decoded invariant to actor (train classifier on four actors, test on fifth held-out actor), or view (train classifier on four views, test on fifth held-out view). Results are each from the average of five different subjects. Error bars represent standard deviation. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with p<0.01 permutation test, and thickness of the line indicates if the significance held for 3, 4, or all 5 subjects.*

Supplemental Figure 3



Supplemental Figure 3: Decoding with eye tracking movement*. We train a classifier on the output of eyetracking data for five subjects as they view five actors perform five actions from five views. Results are from the average of five subjects. Error bars represent standard deviation. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with p<0.01 permutation test. We cannot decoding significantly above chance for five or more consecutive 5ms time bins in any subject, suggesting that the subjects eye movements cannot account for the above decoding performance.*

References

1. Karpathy A, et al. (2014) Large-Scale Video Classification with Convolutional Neural Networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 1725–1732.
2. Le Q V., Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *CVPR 2011* (IEEE), pp 3361–3368.
3. Moeslund TB, Granum E (2001) A Survey of Computer Vision-Based Human Motion Capture. *Comput Vis Image Underst* 81(3):231–268.
4. Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. *Science* 293(5539):2470–3.
5. MICHELS L, LAPPE M, VAINA LM Visual areas involved in the perception of human movement from dynamic form analysis. *Neuroreport* 16(10):1037–1041.
6. Lingnau A, Downing PE (2015) The lateral occipitotemporal cortex in action. *Trends Cogn Sci*. doi:10.1016/j.tics.2015.03.006.
7. Perrett DI, et al. (1985) Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. *Behav Brain Res* 16(2-3):153–170.
8. 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(1):109–129.
9. Vangeneugden J, Pollick F, Vogels R (2009) Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cereb Cortex* 19(3):593–611.
10. Grossman E, et al. (2000) Brain Areas Involved in Perception of Biological Motion. *J Cogn Neurosci* 12(5):711–720.
11. Grossman ED, Blake R (2002) Brain Areas Active during Visual Perception of Biological Motion. *Neuron* 35(6):1167–75.
12. 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(20):11656–61.
13. 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(7):991–1001.
14. Peelen M V, Downing PE (2005) Selectivity for the human body in the fusiform gyrus. *J Neurophysiol* 93(1):603–8.
15. Grossman ED, Jardine NL, Pyles JA (2010) fMR-Adaptation Reveals Invariant Coding of Biological Motion on the Human STS. *Front Hum Neurosci* 4:15.
16. Vangeneugden J, Peelen M V, Tadin D, Battelli L (2014) Distinct neural mechanisms

for body form and body motion discriminations. *J Neurosci* 34(2):574–85.

17. Singer JM, Sheinberg DL (2010) Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *J Neurosci* 30(8):3133–45.
18. Dinstein I, Thomas C, Behrmann M, Heeger DJ (2008) A mirror up to nature. *Curr Biol* 18(1):R13–8.
19. Ogawa K, Inui T (2011) Neural representation of observed actions in the parietal and premotor cortex. *Neuroimage* 56(2):728–35.
20. Wurm MF, Lingnau A (2015) Decoding actions at different levels of abstraction. *J Neurosci* 35(20):7727–35.
21. Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 36(4):193–202.
22. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A* 104(15):6424–9.
23. Mutch J, Knoblich U, Poggio T (2010) CNS: a GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR-2010-013*.
24. Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195(1):215–43.
25. Simoncelli EP, Heeger DJ (1998) A model of neuronal responses in visual area MT. *Vision Res* 38(5):743–761.
26. Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A* 2(2):284–99.
27. Movshon JA, Adelson EH, Gizzi MS, Newsome WT (1986) Pattern Recognition Mechanisms Available at: http://www.cns.nyu.edu/~tony/Publications/movshon-adelson-gizzi-newsome-1985.pdf [Accessed April 21, 2015].
28. Földiák P (2008) Learning Invariance from Transformation Sequences. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.2.194#.VNFu0mTF-14 [Accessed February 4, 2015].
29. Wiskott L, Sejnowski TJ (2006) Slow Feature Analysis: Unsupervised Learning of Invariances. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/089976602317318938#.VNFwQ2TF-14 [Accessed February 4, 2015].
30. Stringer SM, Perry G, Rolls ET, Proske JH (2006) Learning invariant object recognition in the visual system with continuous transformations. *Biol Cybern* 94(2):128–42.
31. Wallis G, Bülthoff HH (2001) Effects of temporal association on recognition memory. *Proc Natl Acad Sci U S A* 98(8):4800–4.
32. Johansson G (1973) Visual perception of biological motion and a model for its analysis. *Percept Psychophys* 14(2):201–211.

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

34. Rolls ET (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27(2):205–18.

35. Rust NC, Dicarlo JJ (2010) Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* 30(39):12978–95.

36. Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object recognition in the human visual system. *J Neurophysiol* 111(1):91–102.

37. Liu H, Agam Y, Madsen JR, Kreiman G (2009) Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62(2):281–90.

38. Carlson T, Hogendoorn H, Fonteijn H, Verstraten FAJ (2011) Spatial coding and invariance in object-selective cortex. *Cortex* 47(1):14–22.

39. Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of object vision: the first 1000 ms. *J Vis* 13(10):1–.

40. Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. *Nat Neurosci* 17(3):455–62.

41. Anselmi F, Leibo JZ, Rosasco L, Tacchetti A, Poggio T (2014) Unsupervised learning of invariant representations with low sample complexity: the magic of sensory cortex or a new framework for machine learning?

42. Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci* 4(3):179–92.

43. Jhuang H, Serre T, Wolf L, Poggio T (2007) A Biologically Inspired System for Action Recognition. *2007 IEEE 11th International Conference on Computer Vision* (IEEE), pp 1–8.

44. Leibo JZ, Mutch J, Poggio TA (2011) Recognition, Why The Brain Separates Face Recognition From Object. *Advances in Neural Information Processing Systems 24*, pp 711–719.

45. 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.

46. Tesche CD, et al. (1995) Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. *Electroencephalogr Clin Neurophysiol* 95(3):189–200.

47. 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(1):212–8.

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

49. Meyers EM (2013) The neural decoding toolbox. *Front Neuroinform* 7. doi:10.3389/fninf.2013.00008.

50. Piccardi M (2004) Background subtraction techniques: a review. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* (IEEE), pp 3099–3104.

51. Tacchetti A, Mallapragada PK, Santoro M, Rosasco L (2013) GURLS: A Least Squares Library for Supervised Learning. *J Mach Learn Res* 14:3201–3205.