PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Invariant recognition drives neural representations of action sequences

Andrea Tacchetti*☯, Leyla Isik☯, Tomaso Poggio

Center for Brains Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA, United States

☯ These authors contributed equally to this work.
* atacchet@mit.edu

OPEN ACCESS

Citation: Tacchetti A, Isik L, Poggio T (2017) Invariant recognition drives neural representations of action sequences. PLoS Comput Biol 13(12): e1005859. https://doi.org/10.1371/journal.pcbi.1005859

Editor: Max Berniker, Northwestern University, UNITED STATES

Received: April 21, 2017

Accepted: October 31, 2017

Published: December 18, 2017

Data Availability Statement: The video sequences used in Experiment 1, 2 and 3 are available for download from The Center for Brains, Minds and Machines: cbmm.mit.edu. The Magnetoencephalography recordings used for our Comparison of model representations and neural recordings are available for download from The Center for Brains, Minds and Machines: cbmm.mit.edu.

Funding: This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award

Abstract

Recognizing the actions of others from visual stimuli is a crucial aspect of human perception that allows individuals to respond to social cues. Humans are able to discriminate between similar actions despite transformations, like changes in viewpoint or actor, that substantially alter the visual appearance of a scene. This ability to generalize across complex transformations is a hallmark of human visual intelligence. Advances in understanding action recognition at the neural level have not always translated into precise accounts of the computational principles underlying what representations of action sequences are constructed by human visual cortex. Here we test the hypothesis that invariant action discrimination might fill this gap. Recently, the study of artificial systems for static object perception has produced models, Convolutional Neural Networks (CNNs), that achieve human level performance in complex discriminative tasks. Within this class, architectures that better support invariant object recognition also produce image representations that better match those implied by human and primate neural data. However, whether these models produce representations of action sequences that support recognition across complex transformations and closely follow neural representations of actions remains unknown. Here we show that spatiotemporal CNNs accurately categorize video stimuli into action classes, and that deliberate model modifications that improve performance on an invariant action recognition task lead to data representations that better match human neural recordings. Our results support our hypothesis that performance on invariant discrimination dictates the neural representations of actions computed in the brain. These results broaden the scope of the invariant recognition framework for understanding visual intelligence from perception of inanimate objects and faces in static images to the study of human perception of action sequences.

Author summary

Recognizing the actions of others from video sequences across changes in viewpoint, gait or illumination is a hallmark of human visual intelligence. A large number of studies have highlighted which areas in the human brain are involved in the processing of biological motion, while others have described how single neurons behave in response to videos of

CCF-1231216. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. We are grateful to the Martinos Imaging Center at MIT, where the neural recordings used in this work were acquired and to the McGovern Institute for Brain Research at MIT for supporting this research. Additional support was provided by the Eugene McDermott Foundation (TP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

human actions. However, little is known about the computational necessities that shaped these neural mechanisms either through evolution or experience. In this paper, we test the hypothesis that this computational goal is the discrimination of action categories from complex video stimuli and across identity-preserving transformations. We show that, within the class of Spatiotemporal Convolutional Neural Networks (ST-CNN), deliberate model modifications leading to representations of videos that better support robust action discrimination, also produce representations that better match human neural data. Importantly, increasing model performance on invariant action recognition leads to a better match with human neural data, despite the model never being exposed to such data. These results suggest that, similarly to what is known for object recognition, supporting invariant discrimination within the constraints of hierarchical ST-CNN architectures drives the neural mechanisms underlying our ability to perceive the actions of others.

Introduction

Humans' ability to recognize the actions of others is a crucial aspect of visual perception. Remarkably, the accuracy with which we can finely discern what others are doing is largely unaffected by transformations that substantially change the visual appearance of a given scene, but do not change the semantics of what we observe (e.g. a change in viewpoint). Recognizing actions, the middle ground between action primitives and activities [1], across these transformations is a hallmark of human visual intelligence, which has proven difficult to replicate in artificial systems. Because of this, invariance to transformations that are orthogonal to a learning task has been the subject of extensive theoretical and empirical investigation in both artificial and biological perception [2,3].

Over the past few decades, artificial systems for action processing have received considerable attention. These methods can be divided into global and local approaches. Some space-time global approaches rely on fitting the present scene to a joint-based model of human bodies, actions are then described as sequences of joint configurations over time [4]. Other global methods use descriptors that are computed using the entire input video at once [5–7]. Local approaches, on the other hand, extract information from video sequences in a bottom-up fashion, by detecting, in their input video, the presence of features that are local in space and time. These local descriptors are then combined, following a hierarchical architecture, to construct more complex representations [8–10]. A specific class of bottom up, local architectures, spatial-temporal Convolutional Neural Networks (ST-CNNs), as well as their recursive extensions [11], are currently the best performing models on action recognition tasks.

Alongside these computational advances, recent studies have furthered our understanding of the neural basis of action perception. Broadly, the neural computations underlying action recognition in visual cortex are organized as a hierarchical succession of spatiotemporal feature detectors of increasing size and complexity [10,12]. In addition, other studies have highlighted of which specific brain areas are involved in the processing of biological motion and actions. In humans and other primates, the Superior Temporal Sulcus, and particularly its posterior portion, is believed to participate in the processing of biological motion and actions [13–20]. In addition to studying which brain regions engage during action processing, a number of studies have characterized the responses of individual neurons. The preferred stimuli of neurons in visual areas V1 and MT are well approximated by moving edge-detection filters and energy-based pooling mechanisms [21,22]. Neurons in the STS region of macaque monkeys respond selectively to actions, are invariant to changes in actors and viewpoint [23] and

their tuning curves are well modeled by simple snippet-matching models [24]. Finally, mirror neurons, cells that exhibit strong responses when subjects are both observing and performing goal directed actions, have been carefully described in recent years [25].

Despite the characterization of the regional and single-unit responses that are involved in constructing neural representations of action sequences, little information is available on what computational tasks might be relevant to explaining and recapitulating how these representations are organized, and in particular which robustness properties are present. The idea of visual representations, internal encodings of incoming stimuli that are useful to the viewer, has a long history in the study of human perception and, since its inception, has provided a powerful tool to link neurophysiology and brain imaging data to more abstract computational concepts like recognition or detection [26–28]. Fueled by advances in computer vision methods for object and scene categorization, recent studies have made progress towards linking neural recordings to computational concepts through quantitatively accurate models of single neurons and entire brain regions. Interestingly, these studies have highlighted a correlation between performance optimization on discriminative object recognition tasks and the accuracy of neural predictions both at the single recording site and neural representation level [29–32]. However, these results have not been extended to action perception and dynamic stimuli.

Here we take advantage of recent advances in artificial systems for action processing to test the hypothesis that invariant recognition drives the representations of action sequences computed by visual cortex. We do so by comparing representations obtained with biologically plausible artificial systems and those measured in human subjects through Magnetoencephalography (MEG) recordings [33]. In this paper we show that, within the Spatiotemporal Convolutional Neural Networks model class [10,12,34,35], deliberate modifications that result in better performing models on invariant action recognition, also lead to empirical dissimilarity matrices that better match those obtained with human neural recordings. Our results suggest that discriminative tasks, and especially those that require generalization across complex transformations, alongside the constraints imposed by the hierarchical organization of visual processing in human cortex, determined which representations of action sequences are computed by visual cortex. Importantly, we quantify the degree of overlap between neural and artificial representations using Representational Similarity Analysis [32]. This measure of agreement between two encodings, does not rely on a one-to-one mapping between neural signal sources and their artificial counterpart, but rather, exploits similarity structures directly in the representation spaces to establish a measure of consensus. Moreover, by highlighting the role of robustness to nuisances that are orthogonal to the discrimination task, our results extend the scope of invariant recognition as a computational framework for understanding human visual intelligence to the study of action recognition from video sequences.

Results

Action discrimination with Spatiotemporal Convolutional representations

We filmed a video dataset showing five actors, performing five actions (drink, eat, jump, run and walk) at five different viewpoints (Fig 1). We then developed four variants of feedforward hierarchical models of visual cortex and used them to extract feature representations of videos showing two different viewpoints, frontal and side. Subsequently, we trained a machine learning classifier to discriminate video sequences into different action classes based on each model's output. We then evaluated the classifier's accuracy in predicting the action content of new, unseen videos.

The four models we developed to extract representations of action sequences from videos were instances of Spatiotemporal Convolutional Neural Networks (ST-CNNs), currently the

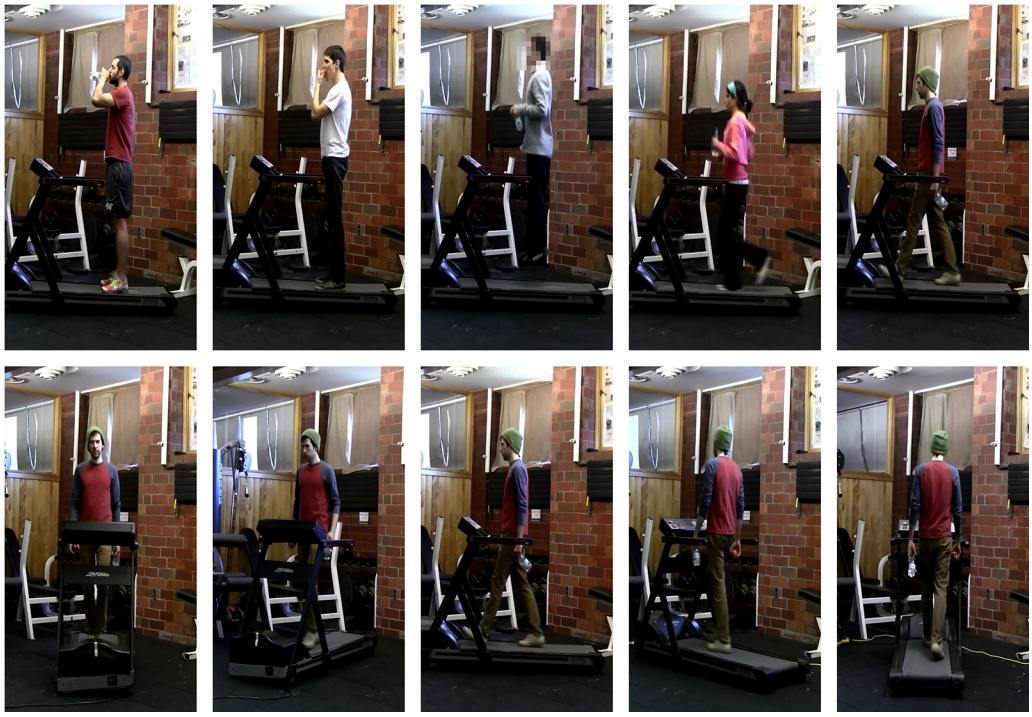

Fig 1. Action recognition stimulus set. Sample frames from action recognition dataset consisting of 2s video clips depicting five actors performing five actions (top row: drink, eat, jump, run and walk). Actions were recorded at five different viewpoints (bottom row: 0-frontal, 45, 90-side, 135 and 180 degrees with respect to the normal to the focal plane), they were all performed on a treadmill and actors held a water bottle and an apple in their hand regardless of the action they performed in order to minimize low-level object/action confounds. Actors were centered in the frame and the background was held constant regardless of viewpoint. The authors who collected the videos identified themselves and the purpose of the videos to the people being video recorded. The individuals agreed to have their videos taken and potentially published.

https://doi.org/10.1371/journal.pcbi.1005859.g001

best performing artificial perception systems for action recognition [34] and were specifically designed to exhibit a varying degree of performance on invariant action recognition tasks. ST-CNN architectures are direct extensions of the Convolutional Neural Networks used to recognize objects or faces in static images [27,36], to input stimuli that extend both in space and time. ST-CNNs are hierarchical models that build selectivity to specific stimuli through template matching operations and robustness to transformations through pooling operations (Fig 2). Qualitatively, Spatiotemporal Convolutional Neural Networks detect the presence of a certain video segment (a template) in their input stimulus; detections for various templates are then aggregated, following a hierarchical architecture, to construct video representations. Nuisances that should not be reflected in the model's output, like changes in position, are discarded through the pooling mechanism [26].

We considered a basic, purely convolutional model, and subsequently introduced modifications to its pooling mechanism and template learning rule to improve performance on invariant action recognition [36]. The first, purely convolutional model, consisted of convolutional layers with fixed templates, interleaved by pooling layers that computed max-operations across contiguous regions of space. In particular, templates in the first convolutional layer contained moving Gabor filters, while templates in the second convolutional layer were sampled from a set of action sequences collected at various viewpoints. The second, Unstructured Pooling model, allowed pooling units in the last layer to span random sets of templates as well as contiguous space regions (Fig 3B). The third, Structured Pooling model, allowed pooling over

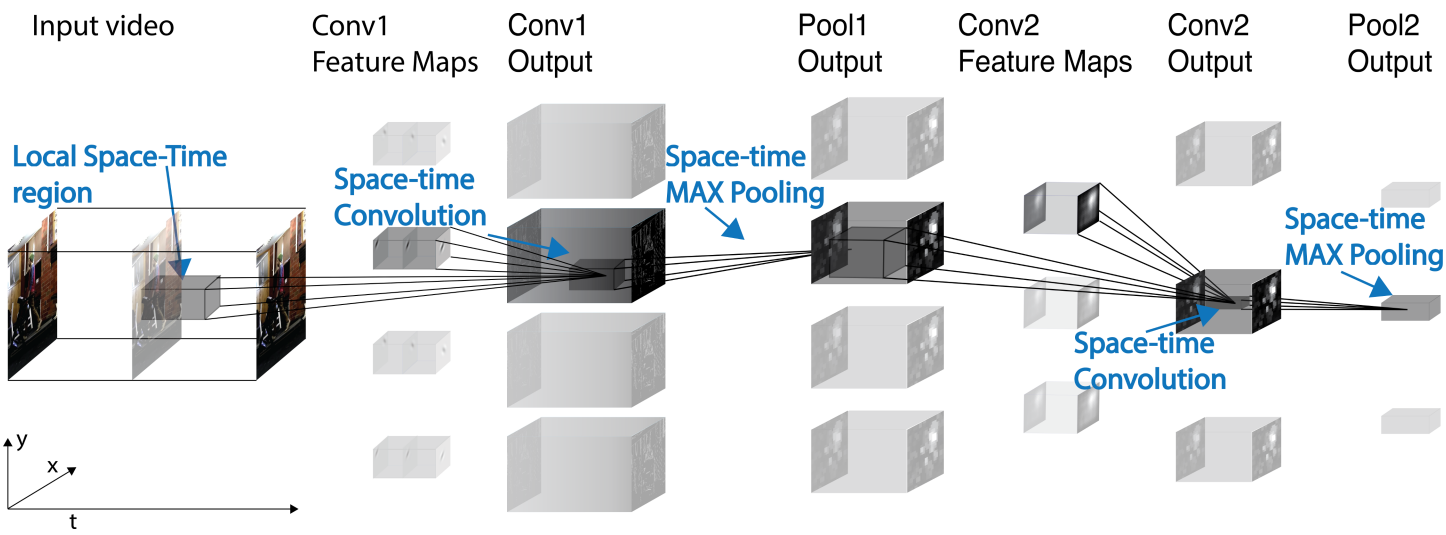

Fig 2. Spatiotemporal Convolutional Neural Networks. Schematic overview of the class of models we used: Spatiotemporal Convolutional Neural Networks (ST-CNNs). ST-CNNs are hierarchical feature extraction architectures. Input videos go through layers of computation and the output of each layer serves as input to the next layer. The output of the last layer constitutes the video representation used in downstream tasks. The models we considered consisted of two convolutional-pooling layers' pairs, denoted as Conv1, Pool1, Conv2 and Pool2. Convolutional layers performed template matching with a shared set of templates at all positions in space and time (spatiotemporal convolution), and pooling layers increased robustness through max-pooling operations. Convolutional layers' templates can be either fixed a priori, sampled or learned. In this example, templates in the first layer Conv1 are fixed and depict moving Gabor-like receptive fields, while templates in the second simple layer Conv2 are sampled from a set of videos containing actions and filmed at different viewpoints. The authors who collected the videos identified themselves and the purpose of the videos to the people being video recorded. The individuals agreed to have their videos taken and potentially published.

https://doi.org/10.1371/journal.pcbi.1005859.g002

contiguous regions of space as well as across templates depicting the same action at various viewpoints. The 3D orientation of each template was discarded through this pooling mechanism, similarly to how position in space is discarded in traditional CNNs (Fig 3A) [2,37]. The fourth and final model employed backpropagation, a gradient based optimization method, to learn convolutional layers' templates by iteratively maximizing performance on an action recognition task [36].

The basic, purely convolutional model we used as a starting point has been shown to be a reliable model of biological motion processing in human visual cortex [10,12]. The modifications we introduced aimed to improve its performance on a challenging invariant action recognition task. In particular, structured and unstructured template pooling mechanisms have been analyzed and theoretically motivated in recent years [2,3]. Moreover, these pooling mechanisms have successfully applied to robust face and object recognition [37]. Finally, backpropagation, the gradient based optimization method used to construct the last model, is widely used in computer vision systems [36], and recently it has been applied to vision science [29,31]. While prima facie this method might not be relevant to brain science (see Discussion), we found here, that the representations obtained with this technique better match human brain data.

We used these models to recognize actions in video sequences in a simple three-steps experimental procedure: first we constructed feedforward hierarchical architectures and used them to extract feature representations of a number of video sequences. We then trained a machine learning classifier to predict the action label of a sequence based on each feature representation. Finally, we quantified the performance of the classifier by measuring prediction accuracy on a set of new unseen videos. The procedure just outlined was performed using three separate subsets of the video dataset described above, one for each step. In particular, constructing spatiotemporal convolutional models requires access to video sequences to sample, or learn,

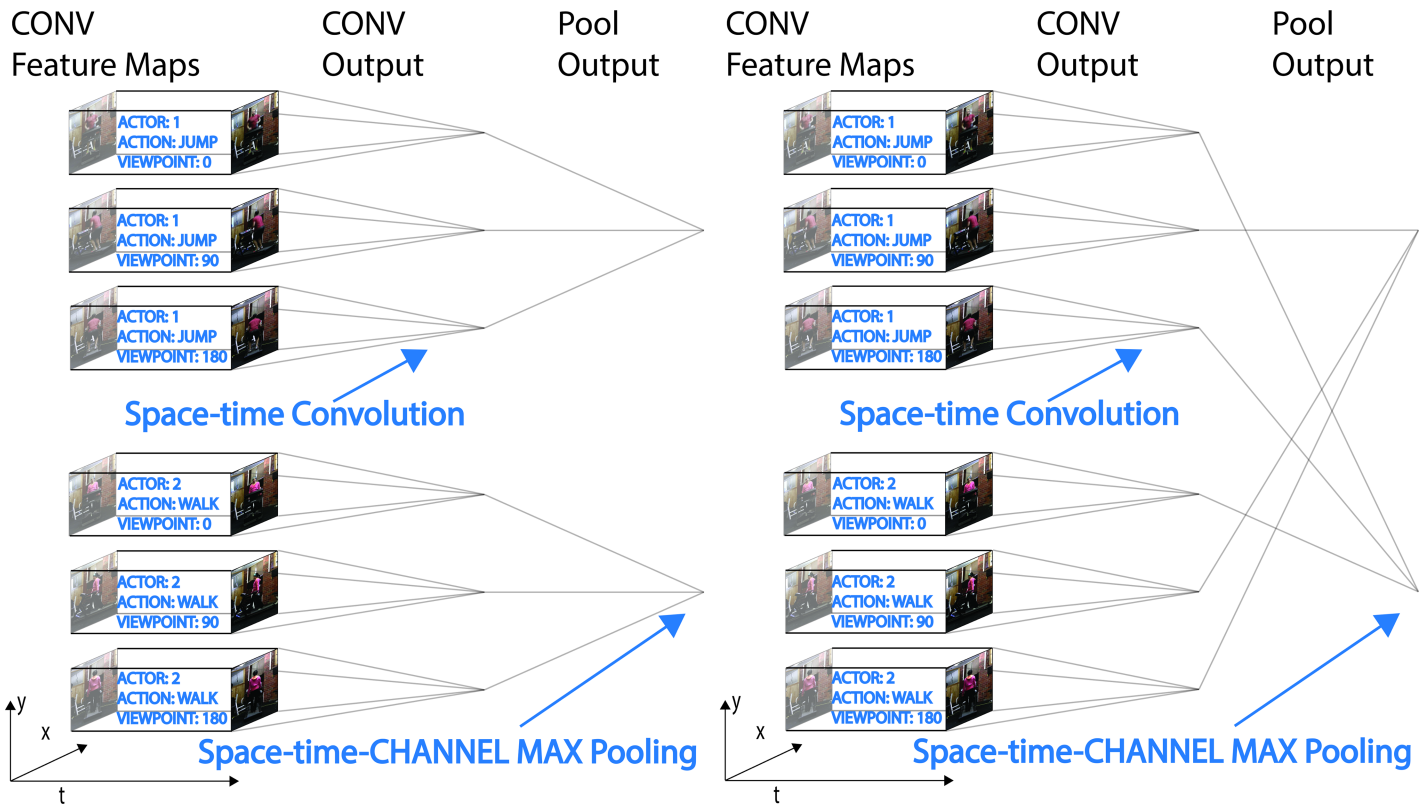

Fig 3. Structured and unstructured pooling. We introduced modifications to the basic ST-CNN to increase robustness to changes in 3D-viewpoint. Qualitatively Spatiotemporal Convolutional Neural Networks detect the presence of a certain video segment (a template) in their input stimulus. The 3D orientation of this template is discarded by the pooling mechanism in our structured pooling model, analogous to how position in space is discarded in a traditional CNN. a) In models with Structured Pooling (model 3, in the main text), the template set for Conv2 layer cells was sampled from a set of videos containing four actors performing five actions at five different viewpoints (see Materials and Methods). All templates sampled from videos of a specific actor and performing a specific action were pooled together by one Pool2 layer unit. b) Models employing Unstructured Pooling (model 2, in the main text) allowed Pool2 cells to pool over the entire spatial extent of their input as well as across channels. These models used the exact same templates employed by models relying on Structured Pooling and matched these models in the number of templates wired to a pooling unit. However, the assignment of templates to pooling was randomized (uniform without replacement) and did not reflect any semantic structure. The authors who collected the videos identified themselves and the purpose of the videos to the people being video recorded. The individuals agreed to have their videos taken and potentially published.

https://doi.org/10.1371/journal.pcbi.1005859.g003

convolutional layers' templates. The subset of videos used for this particular purpose was called the **embedding set**. Likewise, training and testing a classifier requires access to model responses extracted from action sequences; the videos used in these two steps were organized in a **training set** and a **test set**. There was never any overlap between the **test set** and the union of **training** and **embedding set**.

Specifically, we sought to evaluate the four models based on how well they could support discrimination between the five actions in our video dataset both across and within changes in viewpoint. To this end, in Experiment 1, we trained and tested the classifier using model features extracted from videos captured at the same viewpoint while in Experiment 2, we trained and tested the classifier using model features computed from videos at mismatching viewpoints (e.g. if the classifier was trained using videos captured at the frontal viewpoint, then testing would be conducted using videos at the side viewpoint).

Experiment 1: Action discrimination–viewpoint match condition. In Experiment 1, we trained and tested the action classifier using feature representations of videos acquired at the same viewpoint, and therefore did not investigate robustness to changes in viewpoint. In this

case, the **embedding set** contained videos showing all five actions performed at all five viewpoints by four of the five actors. The **training set** was a subset of the **embedding set** and contained all videos at either the frontal or the side viewpoint. Finally, the **test set** contained videos of all five actions performed by the fifth, left-out actor, and performed at the viewpoint matching that shown in the **training set.** All models produced representations that successfully classified videos based on the action they depicted (Fig 4). We observed a significant difference in performance between model 4, the end-to-end trainable model, and fixed template models 1, 2 and 3 (see Methods Section). However, the task considered in Experiment 1 was not sufficient to rank the four types of ST-CNN models.

Experiment 2: Action discrimination–viewpoint mismatch condition. The four ST-CNN models we developed were designed to have varying degrees of tolerance to changes in viewpoint. In Experiment 2, we investigated how well these model representations could support learning to discriminate video sequences based on their action content, across changes in viewpoint. The general experimental procedure was identical to the one outlined for Experiment 1 and used the exact same models. In this case however, we used features extracted from videos acquired at mismatching viewpoints for training and testing (e.g., a classifier trained using videos captured at the frontal viewpoint, would be tested on videos at the side viewpoint). We focused exclusively on to views: 0 and 90 degree with respect to frontal, to test the

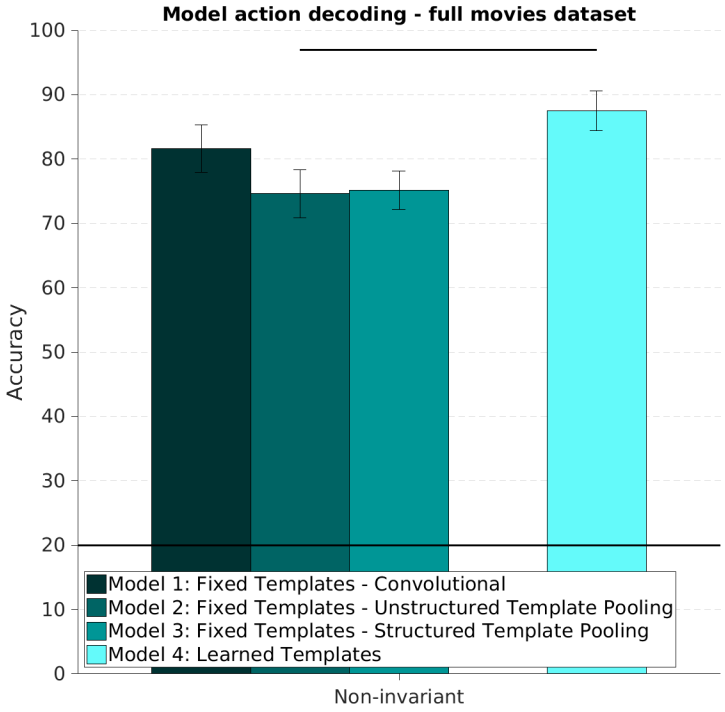

Fig 4. Action recognition: Viewpoint match condition. We trained a supervised machine learning classifier to discriminate videos based on their action content by using the feature representation computed by each of the Spatiotemporal Convolutional Neural Network models we considered. This figure shows the prediction accuracy of a machine learning classifier trained and tested using videos recorded at the same viewpoint. The classifier was trained on videos depicting four actors performing five actions at either the frontal or side view. The machine learning classifier accuracy was then assessed using new, unseen videos of a new, unseen actor performing those same five actions. No generalization across changes in 3D viewpoints was required of the feature extraction and classification system. Here we report the mean and standard error of the classification accuracy over the five possible choices of test actor. Models with learned templates outperform models with fixed templates significantly on this task. Chance is 1/5 and is indicated by a horizontal line. Horizontal lines at the top indicate significant difference between two conditions ($p < 0.05$) based on group ANOVA or Bonferroni corrected paired t-test (see Materials and Methods section).

https://doi.org/10.1371/journal.pcbi.1005859.g004

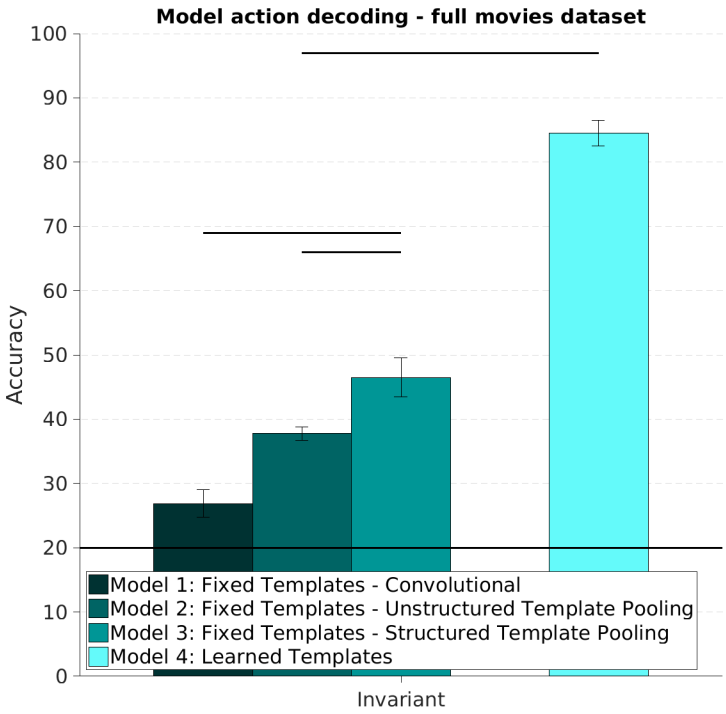

Fig 5. Action recognition: Viewpoint mismatch condition. This figure shows the prediction accuracy of a machine learning classifier trained and tested using feature representations of videos at opposed viewpoints. Hierarchical models were constructed using convolutional templates sampled or learned from videos showing all five viewpoints. During the training and testing of the classifier however, mismatching viewpoints were used. When the classifier was trained using videos at, say, the frontal viewpoint, its accuracy in discriminating new, unseen videos would be established using videos recorded at the side viewpoint. Here we report the mean and standard error of the classification accuracy over the five possible choices of test actor. Models with learned templates resulted in significantly higher accuracy in this task. Among models with fixed templates, Spatiotemporal Convolutional Neural Networks employing Structured pooling outperformed both purely convolutional and Unstructured Pooling models. Chance is 1/5 indicated with horizontal line. Horizontal lines at the top indicate significant difference between two conditions ($p < 0.05$) based on group ANOVA or Bonferroni corrected paired t-test (see Materials and Methods).

https://doi.org/10.1371/journal.pcbi.1005859.g005

same extreme case of generalization across changes in viewpoint (training on a single view that is non-adjacent and non-mirror-symmetric to the test view) as used for the MEG experiments (see Experiment 3 and Materials and Methods). All the models we considered produced representations that were, at least to a minimal degree, useful to discriminate actions invariantly to changes in viewpoint (Fig 5). Unlike what we observed in Experiment 1, it was possible to rank the models we considered based on performance on this task. This was expected, since the various architectures were designed to exhibit various degrees of robustness to changes in viewpoint (see Materials and Methods). The end-to-end trainable models (model 4) performed better than models 1,2 and 3, which used fixed templates, on this task. Within the fixed templates models group, as expected, models that employed a Structured Channel Pooling mechanism to increase robustness performed best [38].

Comparison of model representations and neural recordings

We used Representational Similarity Analysis (RSA) to assess how well each model feature representation, as well as an ideal categorical oracle, matched human neural data. RSA produces a measure of agreement between artificial models and brain recordings based on the correlation between empirical dissimilarity matrices constructed using either the model

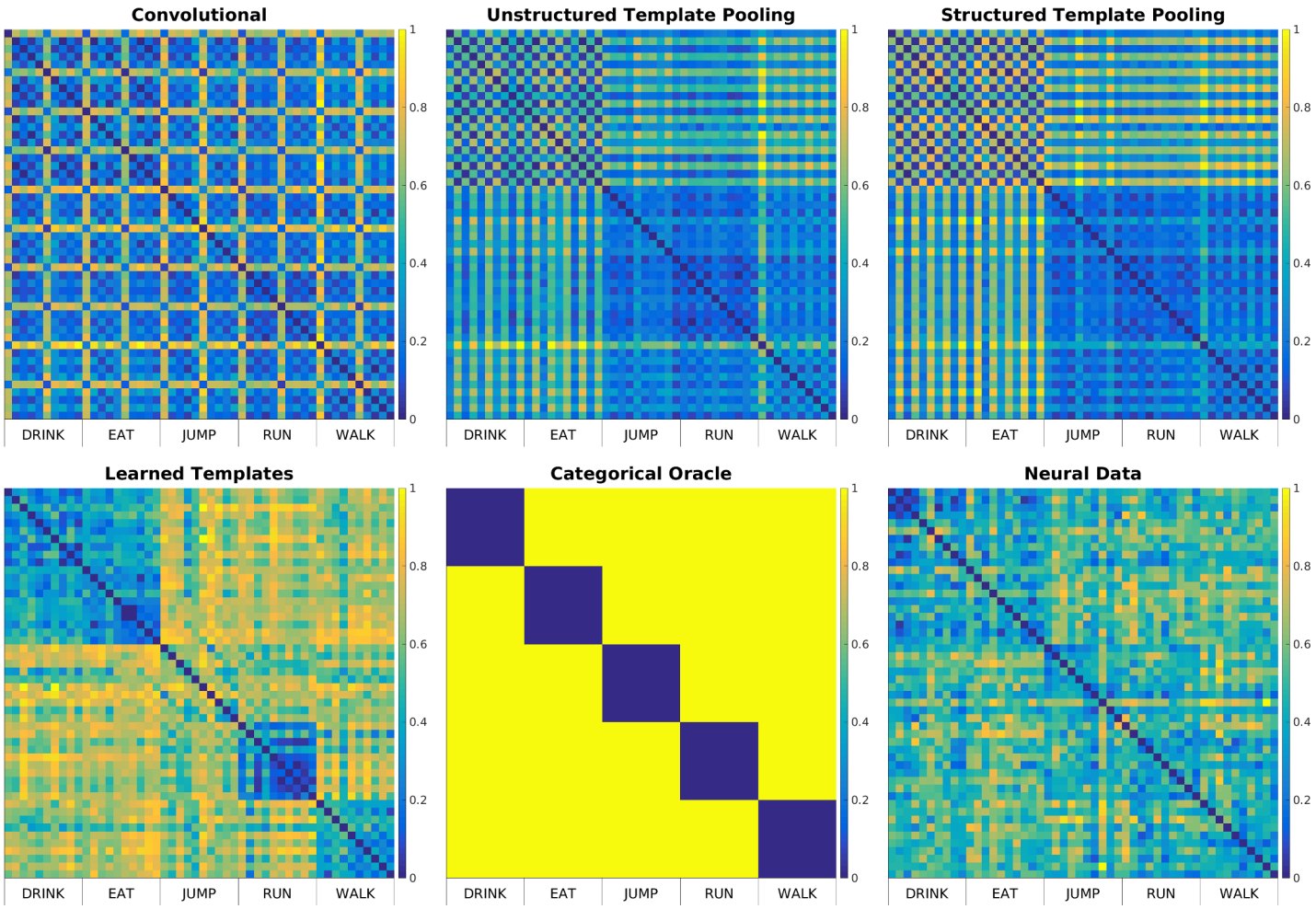

Fig 6. Feature representation empirical dissimilarity matrices. We used feature representations, extracted with the four Spatiotemporal Convolutional Neural Network models, from 50 videos depicting five actors performing five actions at two different viewpoints, frontal and side. Moreover, we obtained Magnetoencephalography (MEG) recordings of human subjects' brain activity while they were watching these same videos, and used these recordings as a proxy for the neural representation of these videos. These videos were not used to construct or learn any of the models. For each of the six representations of each video (four artificial models, a categorical oracle and one neural recordings) we constructed an empirical dissimilarity matrix using linear correlation and normalized it between 0 and 1. Empirical dissimilarity matrices on the same set of stimuli constructed with video representations from a) Model 1: Purely Convolutional model, b) Model 2: Unstructured pooling model, c) Model 3: Structured pooling model d) Model 4: Learned templates model e) Categorical oracle and f) Magnetoencephalography brain recordings.

https://doi.org/10.1371/journal.pcbi.1005859.g006

representation of a set of stimuli, or recordings of the neural responses these stimuli elicit (Fig 6) [32]. We used video feature representations extracted by each model from a set of new, unseen stimuli to construct model dissimilarity matrices. We also constructed dissimilarity matrices using Magnetoencephalograpy (MEG) data from the average of eight subjects viewing the same action video clips. The MEG data consisted of magnetometer and gradiometer recordings from 306 sensors, averaged over a 100ms window centered at the time when action identity was best decoded from these data in a separate experiment [33] (see Materials and Methods). Finally, we constructed a dissimilarity matrix using an action categorical oracle, a simulated ideal observer able to perfectly classify video sequences based on their action content. In this case, the dissimilarity between videos of the same action was zero and the distance across actions was one.

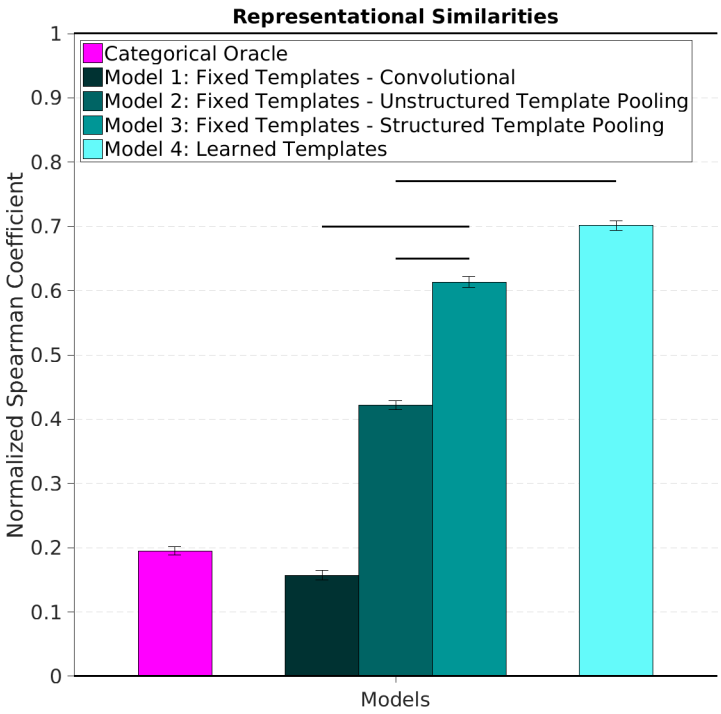

Fig 7. Representational Similarity Analysis between model representations and human neural data. We computed the Spearman Correlation Coefficient (SCC) between the lower triangular portion of the dissimilarity matrix constructed with each of the artificial models we considered and the dissimilarity matrix constructed with neural data (shown and described in Fig 6). We assessed the uncertainty of this measure by resampling the rows and columns of the matrices we constructed. In order to give the SCC score a meaningful interpretation we reported here a normalized score: the SCC is normalized so that the noise ceiling is 1 and the noise floor is 0. The noise ceiling was assessed by computing the SCC between each individual human subjects' dissimilarity matrix and the average dissimilarity matrix over the rest of the subjects. The noise floor was computed by assessing the SCC between the lower portion of the dissimilarity matrix constructed using each of the model representation and a scrambled version of the neural dissimilarity matrix. This normalization embeds the intuition that we cannot expect artificial representations to match human data better than an individual human subject's data matches the mean of other humans and that we should only be concerned care with how much better the models we considered are, on this scale, than a random guess. Models with learned templates agree with the neural data significantly better than models with fixed templates. Among these, models with Structured Pooling outperform both purely Convolutional and Unstructured models. Horizontal lines at the top indicate significant difference between two conditions ($p < 0.05$) based on group ANOVA or Bonferroni corrected paired t-test (see Materials and Methods).

https://doi.org/10.1371/journal.pcbi.1005859.g007

We observed that end-to-end trainable models (model 4) produced dissimilarity structures that better agreed with those constructed from neural data than models with fixed templates (Fig 7). Within models with fixed templates, model 3, constructed using a Structured Pooling mechanism to build invariance to changes in viewpoint, produced representations that agree better with the neural data than models employing Unstructured Pooling (model 2) and purely convolutional models (model 1). The category oracle did not match the MEG data as well as the highest performing models (models 3 and 4), suggesting that improving performance on the action recognition task does not trivially improve matching with the neural data.

Discussion

We have shown that, within the Spatiotemporal Convolutional Neural Networks model class and across a deliberate set of model modifications, feature representations that are more useful to discriminate actions in video sequences in a manner that is robust to changes in viewpoint,

produce empirical dissimilarity structures that are more similar to those constructed using human neural data. These results support our hypothesis that performance on invariant discriminative tasks drives the neural representations of actions that are computed by our visual cortex. Moreover, dissimilarity matrices constructed with ST-CNNs representations match those built with neural data better than a purely categorical dissimilarity matrix. This highlights the importance of both the computational task and the architectural constraints, described in previous accounts of the neural processing of action and motions, to build quantitatively accurate models of neural data representations [39]. Our findings are in agreement with what has been reported for the perception of objects from static images, both at the single recording site and at the whole brain level [29–31], and identify a computational task that explains and recapitulates the properties of the representations of human action in visual cortex.

We developed the four ST-CNN models using deliberate modifications to improve the models' feature representations to invariant action recognition. In so doing, we verified that structured pooling architectures and memory based learning (model 3), as previously described and theoretically motivated [2,3], can be applied to build representations of video sequences that support recognition invariant to complex, non-affine transformations. However, empirically, we found that learning model templates using gradient based methods and a fully supervised action recognition task (model 4), led to better results, both in terms of classification accuracy and agreement with neural recordings [31].

The five actions in our dataset were selected to be highly familiar, include both goal-directed hand-arm movements and whole body movements, and span coarse (run vs. eat) as well as fine (drink vs. eat) action discriminations. While the five actions we considered are far from exhaustive, they allow us rank the performance of our four different models on invariant action recognition. Importantly, we show that our top-performing models capture non-trivial aspects of the neural representations of these actions, as shown by the fact that the ST-CNN models match MEG data better than a categorical oracle.

A limitation of the methods used here is that the extent of the match between a model representation and the neural data is appraised solely based on the correlation between the empirical dissimilarity structures constructed with neural recordings and model representations. This relatively abstract comparison provides no guidance in establishing a one-to-one mapping between model units and brain regions or sub-regions and therefore cannot exclude models on the basis of biological implausibility [30]. In this work, we mitigated this limitation by constraining the model class to reflect previous accounts of neural computational units and mechanisms that are involved in the perception of motion [10,21,22,40,41].

Furthermore, the class of models we developed in our experiments is purely feedforward, however, the neural recordings were maximally action discriminative 470ms after stimulus onset. This late in the visual processing, it is likely that feedback signals are among the energy sources captured by the recordings. These signals are not accounted for in our models. We provide evidence that adding a feedback mechanism, through recursion, does not improve recognition performance nor correlation with the neural data (S1 Fig). We cannot, however, exclude that this is due to the stimuli and discrimination task we designed, which only considered pre-segmented, relatively short action sequences.

Recognizing the actions of others from complex visual stimuli is a crucial aspect of human perception. We investigated the relevance of invariant action discrimination to improving model representations' agreement with neural recordings and showed that it is one of the computational principles shaping the representation of human action sequences human visual cortex evolved, or learned to compute. Our deliberate approach to model design underlined the relevance of both supervised, gradient based, performance optimization methods and memory based, structured pooling methods to the modeling of neural data representations.

While memory-based learning and structured pooling have been investigated extensively as a biologically plausible learning algorithms [2,37,42,43], if and how primate visual cortex could implement gradient based optimization or acquire the necessary supervision remains, despite recent efforts, an unsettled matter [44–46]. Irrespective of the precise biological mechanisms that could carry out performance optimization on invariant discriminative tasks, computational studies point to its relevance to understanding neural representations of visual scenes [29–31]. Recognizing the semantic category of visual stimuli across photometric, geometric or more complex changes, in very low sample regimes is a hallmark of human visual intelligence. By building data representations that support this kind of robust recognition, we have shown here, one obtains empirical dissimilarity structures that match those constructed using human neural data. In the wider context of the study of perception, our results strengthen the claim that the computational goal of human visual cortex is to support invariant recognition by broadening it to the study of action perception.

Materials and methods

Ethics statement

The MIT Committee on the Use of Humans as Experimental Subjects approved the experimental protocol. Subjects provided informed written consent before the experiment. Approval number: 0403000026.

Action recognition dataset

We collected a dataset of five actors performing five actions (drink, eat, jump, run and walk) on a treadmill at five different viewpoints (0, 45, 90, 135 and 180 degrees between the line across the center of the treadmill and the line normal to the focal plane of the video-camera). We rotated the treadmill rather than the camera to keep the background constant across changes in viewpoint (Fig 1). The actors were instructed to hold an apple and a bottle in their hand regardless of the action they were performing, so that objects and background would not differ between actions. Each action/actor/view was filmed for at least 52s. Subsequently the original videos were cut into 26 clips, each 2s long resulting in a dataset of 3,250 video clips. Video clips started at random points in the action cycle (for example a jump might start mid-air or before the actor's feet left the ground) and each 2s clip contained a full action cycle. The authors manually identified one single spatial bounding box that contained the entire body of each actor and cropped all videos according to this bounding box. The authors who collected the videos identified themselves and the purpose of the videos to the people being video recorded. The individuals agreed to have their videos taken and potentially published.

Recognizing actions with spatiotemporal convolutional representations

General experimental procedure. Experiment 1 and Experiment 2 were designed to quantify the amount of action information extracted from video sequences by four computational models of primate visual cortex. In Experiment 1, we tested basic action recognition. In Experiment 2, in particular, we further quantified whether this action information could support action recognition robustly to changes in viewpoint. The motivating idea behind our design is that, if a machine learning classifier is able to discriminate unseen video sequences based on their action content, using the output of a computational model, then this model representation contains some action information. Moreover, if the classifier is able to discriminate videos based on action at new, unseen viewpoints, using model outputs then it must be that these model representations not only carry action information, but that changes in

viewpoint are not reflected in the model output. This procedure is analogous to neural decoding techniques with the important difference that the output of an artificial model is used in lieu of brain recordings [47,48].

The general experimental procedure is as follows: we constructed feedforward hierarchical spatiotemporal convolutional models and used them to extract feature representations of a number of video sequences. We then trained a machine learning classifier to predict the action label of a video sequence based on its feature representation. Finally, we quantified the performance of the classifier, by measuring prediction accuracy on a set of new, unseen videos.

The procedure outlined above was performed using three separate subsets of the action recognition dataset described in the previous section. In particular, constructing spatiotemporal convolutional model requires access to video sequences depicting actions to sample or learn convolutional layers' templates. The subset of video sequences used to learn or sample templates was called **embedding set**. Training and testing the classifier required extracting model responses from a number of video sequences; these sequences were organized in two subsets: **training set** and **test set**. There was never any overlap between the **test set** and the union of **training set** and **embedding set**.

Experiment 1. The purpose of Experiment 1 was to assess how well the data representations produced by each of the four models, supported a non-invariant action recognition task. In particular, the **embedding set** used to sample or learn templates contained videos showing all five actions at all five viewpoints performed by four of the five actors. The **training set** was a subset of the embedding set, and contained videos at either the frontal viewpoint or the side viewpoint. Lastly the **test set** contained videos of all five actions, performed by the fifth left-out actor and performed at either the frontal or side viewpoint. We obtained five different splits by choosing each of the five actors exactly once for test. After the templates had either been learned or sampled we used each model to extract representations of the **train** and **test sets** videos. We averaged the classifier's performance over the two possible choices of training viewpoint, frontal or side. We report the mean and standard error of the classification accuracy across the five possible choices of the test actor.

Experiment 2. Experiment 2 was designed to assess the performance of each model in producing data representations that were useful to classify videos according to their action content, when a generalization across changes in viewpoint was required. The experiment is identical to Experiment 1, and used the exact same models. However, when the **training set** contained videos recorded at the frontal viewpoint, the **test set** would contain videos at side viewpoint and vice-versa. We report the mean and standard deviation over the choice of the test actor of the average accuracy over the choice of training viewpoint.

Feedforward Spatiotemporal Convolutional Neural Networks. Feedforward Spatiotemporal Convolutional Neural Networks (ST-CNNs) are hierarchical models: input video sequences go through layers of computations and the output of each layer serves as input to the next layer (Fig 2). These models are direct generalizations of models of the neural mechanisms that support recognizing objects in static images [26,27], to stimuli that extend in both space and time (i.e. video stimuli). Within each layer, single computational units process a portion of the input video sequence that is compact both in space and time. The outputs of each layer's units are then processed and aggregated by units in the subsequent layers to construct a final signature representation for the whole input video. The sequence of layers we adopted alternates layers of units which perform template matching (or convolutional layers), and layers of units which perform max pooling operations [10,24,34]. Units' receptive field sizes increases as the signal propagates through the hierarchy of layers.

All convolutional units within a layer share the same set of templates (filter bank) and output the dot-product between each filter and their input. Qualitatively, these models work by

detecting the presence of a certain video segment (a template) in the input stimulus. The exact position in space and time of the detection is discarded by the pooling mechanism. The specific models we present here consist of two convolutional-pooling layers' pairs. The layers are denoted as Conv1, Pool1, Conv2 and Pool2 (Fig 2). Convolutional layers are completely characterized by the size, content and stride of their units' receptive fields and pooling layers are completely characterized by the operation they perform (in the cases we considered, output the maximum value of their input) and their pooling regions (which can extend across space, time and filters).

Model 1: Purely convolutional model with sampled templates. The purely convolutional models with fixed and sampled templates we considered were implemented using the Cortical Network Simulator package [49].

The input videos were (128x76 pixel) x 60 frames; the model received the original input videos alongside two scaled-down versions of it (scaling of factors ½ and ¼ in each spatial dimension respectively).

The first layer, Conv1, consisted of convolutional units with 72 templates of size (7x7 pixel) x 3 frames, (9x9 pixel) x 4 frames and (11x11 pixel) x 5 frames. Convolution was carried out with a stride of 1 pixel (no spatial subsampling). Conv1 filters were obtained by letting Gabor-like receptive fields shift in space over frames (as described in previous studies describing the receptive fields of V1 and MT cells [21,22,40]). The full expression for each filter was as follows:

$$G(x, y, t, \theta, \rho, \sigma, \lambda, n) = f(t)\, exp\left(-\frac{(x'(\theta,\rho,t)^2 + y'(\theta,\rho,t)^2)}{2\sigma^2}\right)\cos\left(\frac{2\pi y'}{\lambda}\right)$$

Where $x'(\theta,\rho,t)$ and $y'(\theta,\rho,t)$, are transformed coordinates that take into account a rotation by θ and a shift by ρt in the direction orthogonal to θ. The Gabor filters we considered had a spatial aperture (in both spatial directions) of $\sigma = 0.6\ S$, with S representing the spatial receptive field size and a wavelength $\lambda = \frac{\sqrt{2}}{2}\ \sigma$ [50]. Each filter had a preferred orientation θ chosen among 8 possible orientations (0, 45, 90, 135, 180, 225, 270, 315 degrees with respect to vertical). Each template was obtained by letting the Gabor-like receptive field just described, shift in the orthogonal direction to its preferred orientation (e.g. a vertical edge would move sideways) with a speed ρ chosen from a linear grid of 3 points between 4/3 and 4 pixels per frame (the shift in the first frame of the template was chosen so that the mean of Gabor-like receptive field's envelop would be centered in the middle frame). Lastly, Conv1 templates had time modulation $f(t) = (kt)^2 e^{-kt^2}\left[\frac{1}{n!} - \frac{(kt)^2}{(n+2)!}\right]$ with $n = 3$ and $t = 0,\ldots,T$ with T the temporal receptive field size [10,22].

The second layer, Pool1, performed max pooling operations on its input by simply finding and outputting the maximum value of each pooling region. Responses to each channel in the Conv1 filter bank was pooled independently and units pooled across regions of space: (4x4 units in space) x 1 unit in time with a stride of 2 units in space, and 1 unit in time, and two scale channels. The functional form of the kernel was chosen based on established models of action processing in visual cortex [10].

A second simple layer Conv2, followed Pool1. Templates in this case were sampled randomly from the Pool1 responses to videos in the **embedding set**. We used a model with 512 Conv2 units with sizes (9x9 units in space) x 3 units in time, (17x17 units in space) x 7 units in time and (25x25 units in space) x 11 units in time, and stride of 1 in all directions.

Finally, the Pool2 layer units performed max pooling. Pooling regions extended over the entire spatial input, one temporal unit, all remaining scales, and a single Conv2 channel.

Model 2 and 3: Structured and Unstructured Pooling models with sampled templates. Structured and Unstructured Pooling models (model 2 and 3, respectively) were constructed by modifying the Pool2 layer of the purely convolutional models. Specifically, in these models Pool2 units pooled over the entire spatial input, one temporal unit, all remaining scales, and 9 Conv2 channels, (512 Conv2 channels and 60 Pool2 units mean that some Pool2 units operated on 8 channels and others on 9).

In the models employing a Structured Pooling mechanism, all templates sampled from videos of a particular actor performing a particular action, regardless of viewpoint were pooled together (Fig 3B). Templates of different sizes and corresponding to different scale channels were pooled independently. This resulted in 6 Pool2 units per action/actor pair, one for each receptive-field-size/scale-channel pair. The intuition behind the Structured Pooling mechanism is that the resulting Pool2 units will respond strongly to the presence of a certain template (e.g. the torso of someone running) regardless of its 3D pose [2,37,38,43,51–54].

The models employing an Unstructured Pooling mechanism followed a similar pattern however, the wiring between simple and complex cells was random (Fig 3A). The fixed templates models (model 1,2 and 3) employed the exact same set of templates (we sampled the templates from the embedding sets only once and used them in all three models) and differed only in their pooling mechanisms.

Model 4: Model with learned templates. Models with learned templates were implemented using Torch packages. These models' templates were trained to recognize actions from videos in the embedding set using a Cross Entropy Loss function, full supervision and back-propagation [55]. The models' general architecture was similar to the one we used for models with structured and unstructured pooling. Specifically, during template learning we used two stacked Convolution-BatchNorm-MaxPooling-BatchNorm modules [56] followed by two Linear-ReLU-BatchNorm modules (ReLU units are half-rectifiers) and a final Log-Soft-Max layer. During feature extraction, the Linear and LogSoftMax layers were discarded.

Input videos were resized to (128x76 pixel) x 60 frames, like in the fixed-template models. The first convolutional layer's filter bank comprised 72 filters of size (9x9 pixel) x 9 frames and convolution was applied with stride of 2 in all directions. The first max-pooling layer used pooling regions of size (4x4 units in space) x 1 unit in time and were applied with stride of 2 units in both spatial directions and 1 unit in time. The second convolutional layer's filter bank was made up of 60 templates of size (17x17 units in space) x 3 units in time, responses were computed with a stride of 2 units in time and 1 unit in all spatial directions. The second Max-Pooling layer's units pooled over the full extent of both spatial dimensions, 1 unit in time and 5 channels. Lastly, the Linear layers had 256 and 128 units respectively and free bias terms. Model training was carried out using Stochastic Gradient Descent [55] and mini-batches of 10 videos.

Machine learning classifier. We used the GURLS package [57] to train and test a Regularized Least Squares Gaussian-Kernel classifier using features extracted from the training and test set respectively and the corresponding action labels. The aperture of the Gaussian Kernel as well as the l2 regularization parameter were chosen with a Leave-One-Out cross-validation procedure on the training set. Accuracy was evaluated separately for each class and then averaged over classes.

Significance testing: Model accuracy. We used a group one-way ANOVA to assess the significance of the difference in performance between all the fixed-template methods and the models with learned templates. We then used a paired-sample t-test with Bonferroni correction to assess the significance level of the difference between the performance of individual models. Difference were deemed significant $p < 0.05$.

Quantifying agreement between model representations and neural recordings

Neural recordings. The brain activity of 8 human participants with normal or corrected to normal vision was recorded with an Elekta Neuromag Triux Magnetoencephalography (MEG) scanner while they watched 50 videos (five actors, five actions, two viewpoints: front and side) acquired with the same procedure outlined above, but not included in the dataset used for model template sampling or training. The MEG recordings data was first presented in [33] (the reference also details all acquisition, preprocessing and decoding methods). The MIT Committee on the Use of Humans as Experimental Subjects approved the experimental protocol. Subjects provided informed written consent before the experiment.

In the original neural recording study MEG recordings were used to train a pattern classifier to discriminate video stimuli on the basis of the neural response they elicited. The performance of the pattern classifier was then assessed on a separate set of recordings from the same subjects. This train/test decoding procedure was repeated every 10ms and individually for each subject both in a non-invariant (train and test at the same viewpoint) and an invariant (train at one viewpoint and test at the different viewpoint) case. It was possible to discriminate videos according to their action content based on the neural response they elicited [33].

We used the filtered MEG recordings (all 306 sensors) elicited by each of the 50 videos mentioned above, averaged across subjects and averaged over a 100ms window centered around 470ms after stimulus onset as a proxy to the neural representation of the video (maximum accuracy for action decoding, as reported in the original study, RSA score with the entire time course is shown, for completeness in (S2 Fig).).

Representational Similarity Analysis. We computed the pairwise correlation-based dissimilarity matrix for each of the model representations of the 50 videos that were shown to human subjects in the MEG. Likewise, we computed the empirical dissimilarity matrix computed using MEG neural recordings. We then performed 50 rounds of bootstrap, in each round we randomly sampled 30 videos out of the original 50 (corresponding to 30 rows and columns of the dissimilarity matrices). For each 30-videos sample, we assessed the level of agreement of the dissimilarity matrix induced by each model representation, with the one computed using neural data by calculating the Spearman Correlation Coefficient (SCC) between the lower triangular portions of the two matrices.

We computed an estimate for the noise ceiling in the neural data by repeating the bootstrap procedure outlined above to assess the level of agreement between an individual human subject and the average of the rest. We then selected the highest possible match score across subjects and across 100 rounds of bootstrap to serve as noise ceiling.

Similarly, we assessed a chance level for the Representational Similarity score by computing the match between each model and a scrambled version of the neural data matrix. We performed 100 rounds of bootstrap per model (reshuffling the neural dissimilarity matrix rows and columns each time) and selected the maximum score across rounds of bootstrap and models to serve as baseline score [32].

We normalized the SCC obtained by comparing each model representation to the neural recordings, by re-scaling them to fall between 0 (chance level) and 1 (noise ceiling). In this normalized scale, anything positive matches neural data better than chance with $p < 0.01$.

Significance testing: Matching neural data. We used a one-way group ANOVA to assess the difference between the Spearman Correlation Coefficient (SCC) obtained using models that employed fixed templates and models with learned templates. Subsequently, we assessed the significance of the difference between the SCC of each model by performing a paired t-test

between the samples obtained through the bootstrap procedure. We deemed differences to be significant when $p < 0.05$ (Bonferroni corrected).

Supporting information

S1 Fig. a) Classification accuracy, within and across changes in 3D viewpoint for a Recurrent Convolutional Neural Network. This architecture does not outperform a purely feedforward baseline. b) A Recurrent Convolutional Neural Network does not produce a dissimilarity structure that better agrees with the neural data than a purely feedforward baseline.
(TIF)

S2 Fig. Spearman Correlation Coefficient between the dissimilarity structure constructed using the representation of 50 videos computed from the Spatiotemporal Convolutional Neural Network with learned templates and the neural data over all possible choices of the neural data time bin. Neural data is most informative for action content of the stimulus at the time indicated by the vertical black line [33].
(TIF)

S1 Text. Recurrent neural networks and RSA over time.
(DOCX)

Acknowledgments

We would like to thank Georgios Evangelopoulos, Charles Frogner, Patrick Winston, Gabriel Kreiman, Martin Giese, Charles Jennings, Heuihan Jhuang, and Cheston Tan for their feedback on this work.

Author Contributions

Conceptualization: Andrea Tacchetti, Leyla Isik, Tomaso Poggio.

Data curation: Andrea Tacchetti, Leyla Isik.

Funding acquisition: Tomaso Poggio.

Investigation: Andrea Tacchetti, Leyla Isik.

Methodology: Andrea Tacchetti, Leyla Isik.

Project administration: Tomaso Poggio.

Software: Andrea Tacchetti, Leyla Isik.

Validation: Andrea Tacchetti, Leyla Isik.

Visualization: Andrea Tacchetti, Leyla Isik, Tomaso Poggio.

Writing – original draft: Andrea Tacchetti, Leyla Isik.

Writing – review & editing: Andrea Tacchetti, Leyla Isik, Tomaso Poggio.

References

1. Moeslund TB, Granum E. A Survey of Computer Vision-Based Human Motion Capture. Comput Vis Image Underst. 2001; 81: 231–268. https://doi.org/10.1006/cviu.2000.0897
2. Anselmi F, Leibo JZ, Rosasco L, Mutch J, Tacchetti A, Poggio T. Unsupervised learning of invariant representations. Theor Comput Sci. 2015; http://dx.doi.org/10.1016/j.tcs.2015.06.048
3. Poggio TA, Anselmi F. Visual Cortex and Deep Networks: Learning Invariant Representations. MIT Press; 2016.

4. Johansson G. Visual perception of biological motion and a model for its analysis. Percept Psychophys. 1973; 14: 201–211. https://doi.org/10.3758/BF03212378
5. Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell. 2001; 23: 257–267. https://doi.org/10.1109/34.910878
6. Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst. 2006; 104: 249–257. https://doi.org/10.1016/j.cviu.2006.07.013
7. Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell. 2007; 29: 2247–2253. https://doi.org/10.1109/TPAMI.2007.70711 PMID: 17934233
8. Laptev I. On space-time interest points. International Journal of Computer Vision. 2005. pp. 107–123. https://doi.org/10.1007/s11263-005-1838-7
9. Dollár P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. Proceedings - 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS. 2005. pp. 65–72. https://doi.org/10.1109/VSPETS.2005.1570899
10. Jhuang H, Serre T, Wolf L, Poggio T. A Biologically Inspired System for Action Recognition. IEEE International Conference on Computer Vision (ICCV). 2007. pp. 1–8.
11. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Conf on Computer Vision and Pattern Recognition (CVPR). 2015. pp. 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878
12. Giese MA, Poggio T. Neural Mechanisms for the Recognition of Biological Movements. Nat Rev Neurosci. 2003; 4: 179–192. https://doi.org/10.1038/nrn1057 PMID: 12612631
13. Perrett DI, Smith PA, Mistlin AJ, Chitty AJ, Head AS, Potter DD, et al. Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report. Behav Brain Res. 1985; 16: 153–170. https://doi.org/10.1016/0166-4328(85)90089-0 PMID: 4041214
14. Vangeneugden J, Pollick F, Vogels R. Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. Cereb Cortex. 2009; 19: 593–611. https://doi.org/10.1093/cercor/bhn109 PMID: 18632741
15. Grossman E, Blake R. Brain areas active during visual perception of biological motion. Neuron. 2002; 35: 1167–1175. https://doi.org/10.1016/S0896-6273(02)00897-8 PMID: 12354405
16. Vaina L, Solomon J, Chowdhury S, Sinha P, Belliveau J. Functional neuroanatomy of biological motion perception in humans. Proc Natl Acad Sci. 2001; 98: 11656–61. https://doi.org/10.1073/pnas.191374198 PMID: 11553776
17. Beauchamp M, Lee K, Haxby J, Martin A. FMRI responses to video and point-light displays of moving humans and manipulable objects. J Cogn Neurosci. 2003; 15: 991–1001. https://doi.org/10.1162/089892903770007380 PMID: 14614810
18. Peelen M, Downing P. Selectivity for the human body in the fusiform gyrus. J Neurophysiol. 2005; 93: 603–8. https://doi.org/10.1152/jn.00513.2004 PMID: 15295012
19. Grossman E, Jardine N, Pyles J. fMR-Adaptation Reveals Invariant Coding of Biological Motion on the Human STS. Front Hum Neurosci. 2010; 4: 15. https://doi.org/10.3389/neuro.09.015.2010 PMID: 20431723
20. Vangeneugden J, Peelen M V, Tadin D, Battelli L. Distinct neural mechanisms for body form and body motion discriminations. J Neurosci. 2014; 34: 574–85. https://doi.org/10.1523/JNEUROSCI.4032-13.2014 PMID: 24403156
21. Adelson E, Bergen J. Spatiotemporal energy models for the perception of motion. J Opt Soc Am. 1985; 2: 284–299.
22. Simoncelli E, Heeger D. A model of neuronal responses in visual area MT. Vision Res. 1998; 38: 743–761. PMID: 9604103
23. Singer J, Sheinberg D. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J Neurosci. 2010; 30: 3133–3145. https://doi.org/10.1523/JNEUROSCI.3211-09.2010 PMID: 20181610
24. Tan C, Singer J, Serre T, Sheinberg D, Poggio T. Neural representation of action sequences: how far can a simple snippet-matching model take us? Adv Neural Inf Process Syst. 2013; 593–601.
25. Kilner JM, Lemon RN. What we know currently about mirror neurons. Curr Biol. 2013; 23. https://doi.org/10.1016/j.cub.2013.10.051 PMID: 24309286
26. Serre T, Oliva A, Poggio T. A feedforward architechture accounts for rapid categorization. Proc Natl Acad Sci. 2007; 104: 6424–6429. https://doi.org/10.1073/pnas.0700622104 PMID: 17404214
27. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999; 2: 1019–1025. https://doi.org/10.1038/14819 PMID: 10526343

28. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980; 36: 193–202. PMID: 7370364

29. Yamins D, Hong H, Cadieu C, Solomon E, Seibert D, DiCarlo J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci. 2014; 111: 8619–24. https://doi.org/10.1073/pnas.1403112111 PMID: 24812127

30. Yamins D, DiCarlo J. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 2016; 19: 356–365. https://doi.org/10.1038/nn.4244 PMID: 26906502

31. Khaligh-Razavi SM, Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol. 2014; 10. https://doi.org/10.1371/journal.pcbi.1003915 PMID: 25375136

32. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2008; https://doi.org/10.3389/neuro.06.004.2008 PMID: 19104670

33. Isik L, Tacchetti A, Poggio TA. A fast, invariant representation for human action in the visual system. J Neurophysiol. American Physiological Society; 2017; https://doi.org/10.1152/jn.00642.2017 PMID: 29118198

34. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li FF. Large-scale video classification with convolutional neural networks. IEEE Conf on Computer Vision and Pattern Recognition (CVPR). 2014. pp. 1725–1732. https://doi.org/10.1109/CVPR.2014.223

35. Ji S, Yang M, Yu K, Xu W. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2013; 35: 221–31. https://doi.org/10.1109/TPAMI.2012.59 PMID: 22392705

36. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521: 436–444. https://doi.org/10.1038/nature14539 PMID: 26017442

37. Leibo J, Liao Q, Anselmi F, Poggio TA. The Invariance Hypothesis Implies Domain-Specific Regions in Visual Cortex. PLoS Comput Biol. Public Library of Science; 2015; 11. https://doi.org/10.1371/journal.pcbi.1004390 PMID: 26496457

38. Leibo J, Mutch J, Poggio T. Learning to discount transformations as the computational goal of visual cortex? Present FGVC/CVPR 2011, Color Springs, CO. 2011;

39. Jhuang H. A biologically inspired system for action recognition. 2008; 58.

40. Rust N, Schwartz O, Movshon JA, Simoncelli E. Spatiotemporal elements of macaque v1 receptive fields. Neuron. 2005; 46: 945–956. https://doi.org/10.1016/j.neuron.2005.05.021 PMID: 15953422

41. Rust N, Mante V, Simoncelli E, Movshon JA. How MT cells analyze the motion of visual patterns. Nat Neurosci. 2006; 9: 1421–1431. https://doi.org/10.1038/nn1786 PMID: 17041595

42. Leibo J, Liao Q, Anselmi F, Freiwald W, Poggio T. View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation. Curr Biol. 2016; 27: 1–6. https://doi.org/10.1016/j.cub.2016.10.044 PMID: 27916526

43. Wiskott L, Sejnowski T. Slow feature analysis: Unsupervised learning of invariances. Neural Comput. 2002; 14: 715–770. https://doi.org/10.1162/089976602317318938 PMID: 11936959

44. Bengio Y, Lee D-H, Bornschein J, Lin Z. Towards Biologically Plausible Deep Learning. arxiv:15020415. 2015; 1–18. https://doi.org/10.1007/s13398-014-0173-7.2

45. Liao Q, Leibo J, Poggio T. How Important is Weight Symmetry in Backpropagation? arXiv:151005067. 2015;

46. Mazzoni P, Andersen R, Jordan M. A more biologically plausible learning rule for neural networks. Proc Natl Acad Sci. 1991; 88: 4433–4437. https://doi.org/10.1073/pnas.88.10.4433 PMID: 1903542

47. Jacobs A, Fridman G, Douglas R, Alam N, Latham P, Prusky G, et al. Ruling out and ruling in neural codes. Proc Natl Acad Sci. National Acad Sciences; 2009; 106: 5936–5941. https://doi.org/10.1073/pnas.0900573106 PMID: 19297621

48. Isik L, Meyers E, Leibo J, Poggio T. The dynamics of invariant object recognition in the human visual system. J Neurophysiol. 2014; 111: 91–102. https://doi.org/10.1152/jn.00394.2013 PMID: 24089402

49. Mutch J, Knoblich U, Poggio T. CNS: a GPU-based framework for simulating cortically-organized networks. MIT-CSAIL-TR. 2010;2010–13.

50. Mutch J, Anselmi F, Tacchetti A, Rosasco L, Leibo J, Poggio T. Invariant Recognition Predicts Tuning of Neurons in Sensory Cortex. Computational and Cognitive Neuroscience of Vision. Singapore: Springer Singapore; 2017. pp. 85–104.

51. Liao Q, Leibo J, Poggio T. Unsupervised learning of clutter-resistant visual representations from natural videos. arXiv:14093879. 2014;

52. Stringer SM, Perry G, Rolls ET, Proske JH. Learning invariant object recognition in the visual system with continuous transformations. Biol Cybern. 2006; 94: 128–142. https://doi.org/10.1007/s00422-005-0030-z PMID: 16369795

53. Wallis G, Bülthoff H. Effects of temporal association on recognition memory. Proc Natl Acad Sci. 2001; 98: 4800–4804. https://doi.org/10.1073/pnas.071028598 PMID: 11287633

54. Földiák P. Learning Invariance from Transformation Sequences. Neural Comput. 1991; 3: 194–200. https://doi.org/10.1162/neco.1991.3.2.194

55. Kingma D, Ba JL. Adam: a Method for Stochastic Optimization. Int Conf Learn Represent. 2015; 1–13.

56. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:150203167. 2015; 1–11. https://doi.org/10.1007/s13398-014-0173-7.2

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