Neuropsychologia 127 (2019) 35–47

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling

Burcu A. Urgen[a,b,c,*,1], Selen Pehlivan[d,1], Ayse P. Saygin[e,f]

[a] Department of Psychology, Bilkent University, Ankara, Turkey
[b] National Magnetic Resonance Research Center and Aysel Sabuncu Brain Research Center, Bilkent University, Ankara, Turkey
[c] Graduate School of Science and Engineering, Interdisciplinary Neuroscience Program, Bilkent University, Ankara, Turkey
[d] Department of Computer Engineering, TED University, Ankara, Turkey
[e] Department of Cognitive Science, UC San Diego, La Jolla, CA, USA
[f] Neurosciences Program, UC San Diego, La Jolla, CA, USA

ARTICLE INFO

Keywords:
Action perception
Computer vision
FMRI
Representational similarity analysis
Modeling
PSTS
Inferior parietal cortex
Ventral premotor cortex

ABSTRACT

Visual processing of actions is supported by a network consisting of occipito-temporal, parietal, and premotor regions in the human brain, known as the Action Observation Network (AON). In the present study, we investigate what aspects of visually perceived actions are represented in this network using fMRI and computational modeling. Human subjects performed an action perception task during scanning. We characterized the different aspects of the stimuli starting from purely visual properties such as form and motion to higher-aspects such as intention using computer vision and categorical modeling. We then linked the models of the stimuli to the three nodes of the AON with representational similarity analysis. Our results show that different nodes of the network represent different aspects of actions. While occipito-temporal cortex performs visual analysis of actions by means of integrating form and motion information, parietal cortex builds on these visual representations and transforms them into more abstract and semantic representations coding target of the action, action type and intention. Taken together, these results shed light on the neuro-computational mechanisms that support visual perception of actions and provide support that AON is a hierarchical system in which increasing levels of the cortex code increasingly complex features.

1. Introduction

In our daily life, in a lecture room, at grocery shopping, in traffic, or at work, we constantly observe other people in action. This simple skill, action perception, is very important for survival because it allows us to take the appropriate action based on what we see. For instance, if somebody waves at you, you probably want to smile and wave back. On the other hand, if somebody attempts to attack you, you probably want to run away.

Due to its evolutionary importance, the neural systems that support action perception have been an intense area of research in neuroscience. There are two lines of research that focus on different aspects of actions. The first line of research, which we refer as visual neuroscience of action perception, focuses on early stages of visual processing and how actions are processed in the early visual cortex (Giese and Poggio, 2003; Blake and Shiffrar, 2007). Two important visual cues in an observed action are the form and motion of the actor. Accordingly, visual processing of actions has been studied within the framework of two parallel pathways of the visual system (Mishkin and Ungerleider, 1982): a dorsal pathway which primarily process motion information, and a ventral pathway which primarily process form information. The question for action perception is how these form and motion information are integrated together to give the percept of a moving entity.

On the other hand, a second line of research, which we refer as cognitive neuroscience of action perception, focuses on brain regions that process actions beyond visual cortex. This line of research has identified a network, commonly known as the Action Observation Network, consisting of three core regions in posterior superior temporal sulcus (pSTS), parietal, and premotor cortex (Rizzolatti and Craighero, 2004; Cross et al., 2008; Caspers et al., 2010; Nelissen et al., 2011). This network is hypothesized to be a hierarchical system (Kilner et al., 2007) in which pSTS gets the visual form and motion information from the visual cortex and communicate them to the parietal and premotor

* Corresponding author at: Department of Psychology, Bilkent University, Ankara, Turkey.
E-mail addresses: burcu.urgen@bilkent.edu.tr (B.A. Urgen), selen.pehlivan@tedu.edu.tr (S. Pehlivan), asaygin@cogsci.ucsd.edu (A.P. Saygin).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.neuropsychologia.2019.02.006
Received 28 November 2018; Received in revised form 4 February 2019; Accepted 11 February 2019
Available online 14 February 2019
0028-3932/

cortex which process more complex aspects of actions.

The aim of the present study is to link the research in visual neuroscience and cognitive neuroscience of action perception with a computational approach to improve our understanding of how actions are processed in the human brain at a systems level. Researchers who study the visual neuroscience of action perception have understandably used simple stimuli (e.g. point-light displays instead of ecologically valid action videos) to target the form and motion processing. On the other hand, researchers who study the cognitive neuroscience of action perception have used relatively complex action videos to target some higher-level aspects of actions such as the actor or the goal without controlling the low-level form and motion aspects. Therefore, it has not yet been possible to investigate how different aspects of actions, including form and motion as well as high-level aspects, are coded in the human brain using complex, ecologically valid but to a certain extent controlled action videos in a single study. The present study aims to fill this gap in knowledge and bring rigor to the field by introducing computational characterization of complex action stimuli and how they link to human brain responses. Given that we have a relatively better understanding of the visual system we focus on the Action Observation Network but link it with the visual system.

With the advances in multivariate pattern analysis methods in cognitive neuroscience, one approach to understand what aspects of actions the human brain represents is to computationally model the different aspects of the stimuli and relate these models to brain responses. This approach has been fruitful in various domains of cognitive neuroscience such as in understanding the representations in early visual cortex and inferior-temporal cortex during object recognition (Clarke and Tyler, 2014; Khaligh-Razavi and Kriegeskorte, 2014; Jozwik et al., 2016, 2017). Actions are complex stimuli and could be modeled in various different ways to understand the neural representations in different nodes of the AON. In the present study, we take such a computational approach and model the action stimuli we use in an action perception study. We then link the models of the stimuli to fMRI responses using representational similarity analysis. The models come in two types: computer vision models that characterize visual form and motion of the stimuli with varying degrees of complexity, and categorical models that capture high-level visual and semantic aspects of the stimuli. We investigate what aspects of actions are represented in the Action Observation Network.

2. Materials and methods

2.1. Participants

27 subjects (12 females, 15 males) from the undergraduate and graduate student community at the University of California, San Diego participated in the study. Data of 4 subjects were not included in the data analysis due to large head movements (3 subjects) and technical problems in data acquisition (1 subject). The reported results included 18 subjects as all the ROIs under investigation were identified in 18 of them (see Section 3.1). The subjects had normal or corrected-to-normal vision and no history of neurological disorders. Informed consent was obtained in accordance with UCSD Human Research Protections Program. The subjects were paid $25 for 1.5 hours participation in the study.

2.2. Stimuli

Stimuli were video clips of actions performed by 3 agents: the humanoid robot Repliee Q2 in two different appearances (robotic and human-like appearances) and by the human 'master' after whom Repliee Q2 was modeled. We call these agents Robot (Agent 1), Android (Agent 2), and Human (Agent 3), respectively (Fig. 1; also see (Saygin et al., 2012; Urgen et al., 2013) for additional details about the stimuli). The robot's movement kinematics was mechanical differing from dynamics of biological motion. All the agents performed 8 different actions. The actions included drinking from a cup, grasping an object, handwaving, talking (for introducing herself), nudging, throwing a piece of paper, turning to the right, wiping a table.

2.3. Procedure

Since prior knowledge can induce cognitive biases against artificial agents (Saygin and Cicekli, 2002), each participant was given exactly the same introduction to the study and same exposure to the stimuli. Before starting fMRI scans, subjects were shown each video and were told whether each agent was a human or a robot (and thus were not uncertain about the identity of the agents during the experiment).

We recorded fMRI BOLD response as subjects watched 2 s video clips of the three agents performing eight different body actions (drinking from a cup, grasping an object, handwaving, talking, nudging, throwing a paper, turning to right, wiping a table). Each subject was scanned for 8 runs in one session. In each run, the experiment had a block design in which blocks consisted of video clips of one agent type (Human, Android, or Robot, see Fig. 1). The experiment had 18 stimuli blocks (6 Human, 6 Android, 6 Robot) and they were presented in a pseudo-randomized order ensuring that all order combinations were presented (i.e. H-A-R, H-R-A, A-H-R, A-R-H, R-H-A, R-A-H). Presentation of three blocks of the agents was always followed by a rest block in which subjects fixated a cross for a time interval varying between 8.1 s and 13.5 s. Each block had 9 trials (8 different actions and repetition of a randomly chosen action once) with 0.1 s inter-stimulus interval in between the trials. Each subject was presented a different order of blocks and of stimuli within each block. Subjects performed a 1-back task throughout the experiment by pressing a button whenever a movie was repeated in a block.

2.4. fMRI image acquisition and preprocessing

We scanned our subjects at the Center for fMRI at University of California, San Diego using the 3T GE MR750 scanner (TR = 2700 ms, TE = 30, Flip angle = 90, number of slices = 35, voxel size = 3 mm × 3 mm × 3 mm, 152 volumes in each run, sequential acquisition). The subjects viewed the stimuli presented on a projector through a mirror mounted on the head cover in the scanner. After scanning, the fMRI data of each subject were pre-processed with standard procedures including motion correction, slice-time correction, normalization, and smoothing using SPM8.[2] Then, two different first-level analyses were done using general linear model (GLM). In the first analysis, each agent type (Human, Android, Robot) as well as the rest blocks (fixation) were modeled as a separate condition and beta images were generated for these conditions. This analysis was done to identify the overall activity patterns and determine the ROIs of the AON. In the second analysis, each trial of all stimulus types was modeled separately and beta images corresponding to each trial were generated for each voxel. This analysis was done to prepare the single trials for RSA. Motion parameters generated in the preprocessing stage were used as regressors in both analyses.

2.5. Identification of Region of Interests (ROIs)

We identified the three ROIs of the AON, pSTS, inferior parietal lobe, and ventral premotor cortex by contrasting the overall activation patterns for all stimulus conditions compared to fixation ($p < 0.001$ uncorrected) using the first first-level analysis for each subject (described in Section 2.4 above). Then, we chose the central voxel of the activation in pSTS, inferior parietal, and ventral premotor cortex for each subject that was normalized to common template, and extracted a

[2] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/.

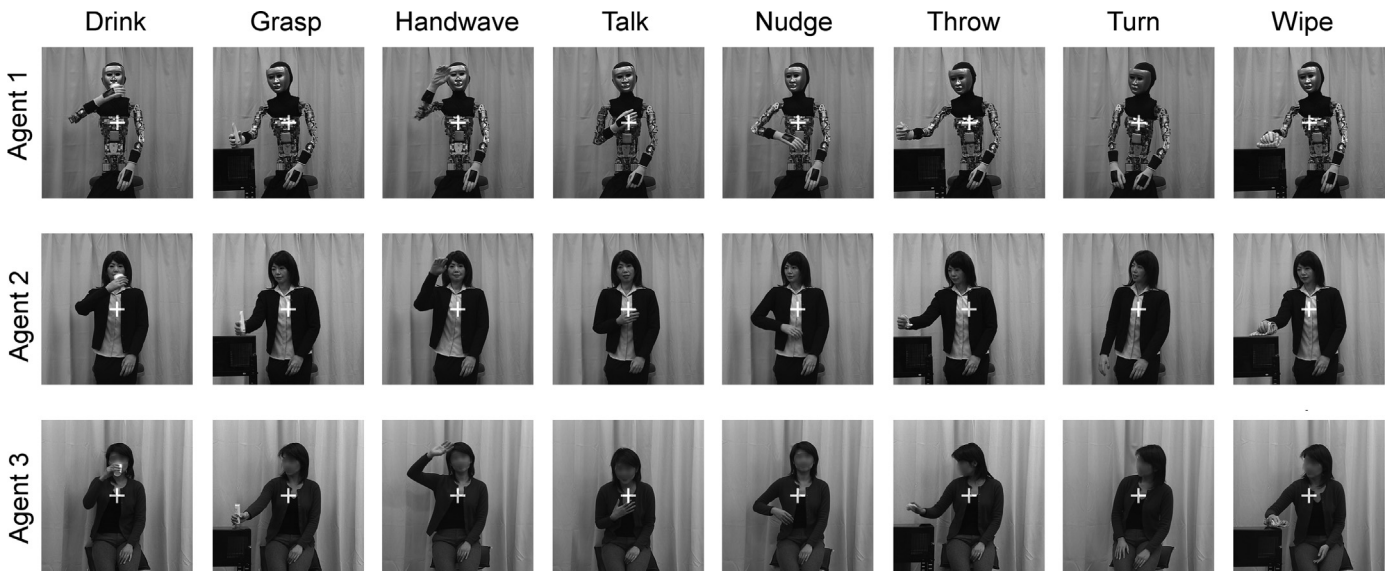

Fig. 1. Still frames from the 24 videos in the study. There were 3 agents (one mechanical robot (Agent 1), one human-like robot (Agent 2), and one human (Agent 3)) and 8 actions (drink, grasp, handwave, talk, nudge, throw, turn, wipe). The authors who took the photograph identified themselves and the purpose of the photograph to the people being photographed in the figure, and the individuals agreed to have their photograph taken and published. The face of the human agent is blurred here for publication purposes, it was visible in the experiment.

sphere ROI with 4 mm radius that covers the activation pattern in each subject. The radius of the sphere was determined by examining the activation of single subjects in each ROI. Although the activation in some ROIs were more extensive in some subjects, we stayed conservative and determined the smallest ROI size that can be identified in common in all subjects, which was 4 mm.

2.6. Data analysis: Representational similarity analysis (RSA)

In order to investigate and dissociate the representational content of the core nodes of the AON, namely pSTS, parietal, and premotor nodes, we used representational similarity analysis and linked the brain responses to that of two sets of models. The first model set includes computer vision models that have been successfully used in action recognition research. We hypothesized that pSTS, the region of AON that is lowest in the hierarchy and closest to the visual areas lower in the hierarchy (such as MT) would have significant and better correlations with computer vision models than the inferior parietal and ventral premotor nodes. We were further interested to find out which computer vision models would best represent pSTS since this would allow us to have insights about the computations pSTS carries out during action perception. The second model set includes attribute-based categorical models that model visual or semantic aspects of the action videos. We hypothesized that the parietal node would have significant and better correlations with attribute-based categorical models as compared to pSTS and premotor nodes since it lies higher in the cortical hierarchy in the AON and it would be an area where visual information about actions is transformed into a more semantic representation.

The RDM of each brain ROI were linked to the RDM of each model by taking their Kendall tau correlation distance (1− Kendall tau) (Nili et al., 2014) excluding the diagonals (Ritchie et al., 2017). This gives us a correlation coefficient which indicates how strong the brain and model patterns are, and a *p*-value which indicates the statistical significance of the correlation, which we used in the evaluation of multiple models for a given ROI. We corrected for multiple comparisons using FDR (Benjamini and Hochberg, 1995) for the number of models and ROIs (we first considered $p < 0.05$ for the models, and then also corrected for the 3 ROIs resulting in $0.05/3 = p < 0.016$). In addition to the statistical significance of the model-brain correlations, we also computed the upper and lower bounds of noise ceiling for model-brain correlations as implemented in RSA toolbox (Nili et al., 2014) for the model evaluations. Any model that exceeded the lower bound of the noise ceiling and that had significant correlation with the brain was considered to be a candidate model for that ROI's representation.

2.6.1. Brain RDMs: pSTS, parietal, and premotor nodes

We calculated representational dissimilarity matrices (RDMs) in pSTS, parietal, and premotor nodes for each subject by taking the correlation distance between all pairs for stimuli using the beta images derived in the first-level analysis in SPM8, which resulted in a 24×24 matrix. We computed the grand average similarity matrix by taking the average of all subjects for each ROI. All the steps in RSA were performed with custom scripts in MATLAB.[3]

2.6.2. Computer vision models

We constructed the $N \times N$ RDMs for the video stimuli using several computer vision models (Herath et al., 2017) that have been successfully used in action recognition (N corresponds to 24 samples of our video stimuli). Particularly, each video was represented as a feature vector output from one of these models, and pair-wise dissimilarities were computed by taking the Spearman correlation for each pair of video representations to construct the computer vision model RDMs.

We examined five models of computer vision on video stimuli. The motivation for including these particular models was to be able to model two basic visual features of action stimuli, namely form and motion, and their combination with varying levels of complexity. The first three models, including gabor, optical flow and trajectory models, utilized video descriptors with simple design principles to emphasize the importance of these features for low-level visual understanding (Gabor modeling form, optical flow modeling motion, and trajectory model modeling low-level form and motion). Remaining two models, namely space-time interest points and dense trajectory models, were selected from the action recognition literature as robust video descriptors with bag-of-features (BoF) representation (Fei-Fei and Perona, 2005). Particularly, both models capture the local information of appearance and the local information of motion, and represent a higher-level model as compared to trajectory model (KLT below). Below, we

[3] http://www.mathworks.com/products/matlab/.

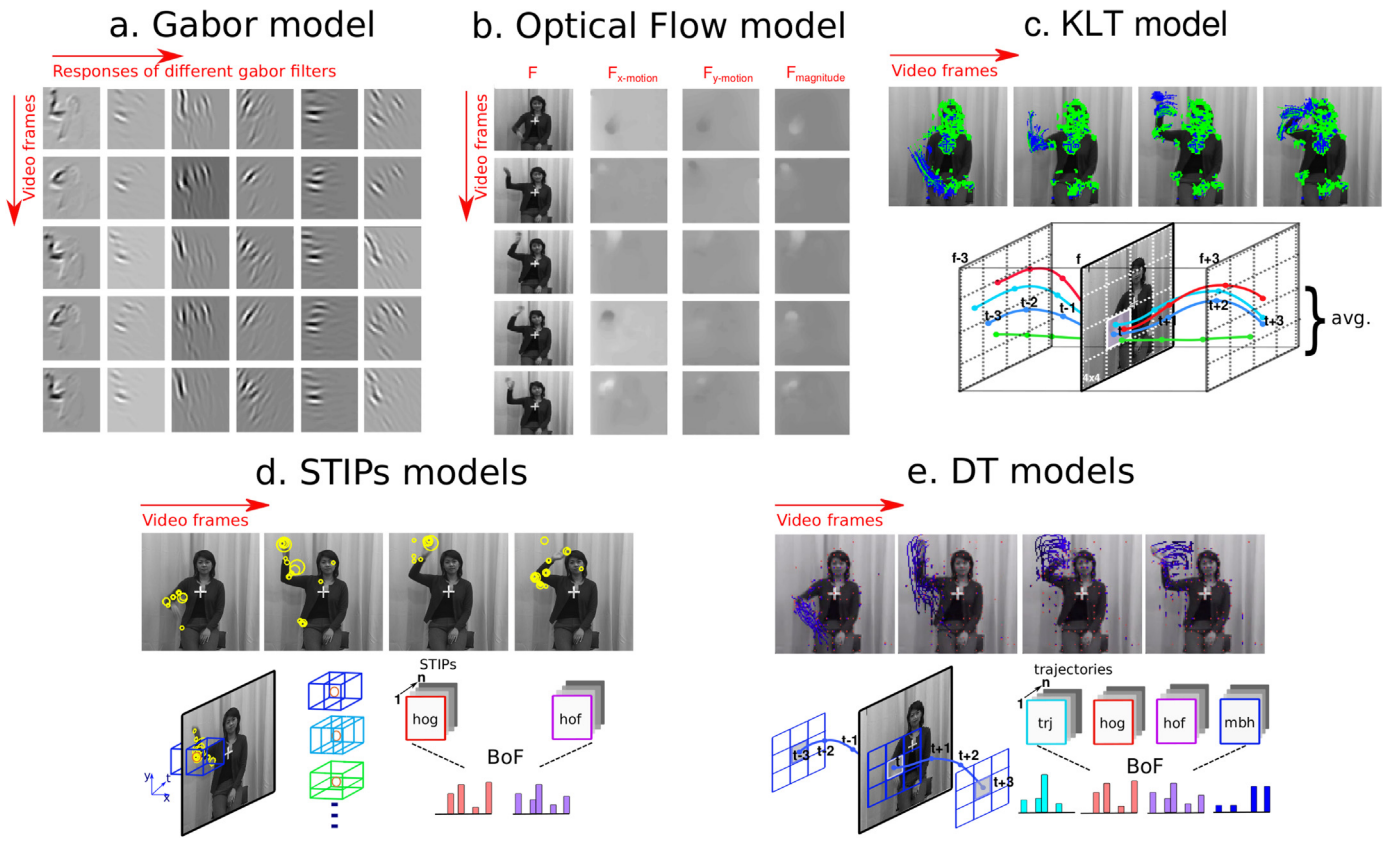

Fig. 2. Selected models of computer vision to represent video stimuli. (a) Gabor Model, (b) Optical Flow model, (c) Kanade-Lucas-Tomasi (KLT) Trajectory model, (d) Space-time Interest Points (STIPs) model with STIP-hog and STIP-hof, and (e) Dense Trajectory (DT) model with DT-trj, DT-hog, DT-hof and DT-mbh.

provide the details of the computer vision models and how the video stimuli were represented as descriptors of these models (see Fig. 2 for the computer vision models). For all the models, the videos of the stimuli were extracted with frames (~ 58 grayscale frames) of size 96 × 96 as inputs.

Gabor model The Gabor video model is the representation of video stimulus using gabor filter activations. First, a bank of 2D-gabor filters was constructed to capture the variations in different orientations, scales and frequencies (Olshausen et al., 1996), where our bank consisted of 144 spatial 2D-Gabor filters in 8 orientations, 9 scales and 2 frequencies. Next, each frame of every video stimulus was filtered using filters of the gabor bank and the activation frames were downsampled to half resolution by skipping two pixels. The Gabor descriptor of the video was the stacked activations of all frames.

Optic Flow model The Optical Flow model is the representation of motion pattern on frames of video stimuli. Optical flow based models are less variant to appearance with strong motion clues. Inspiring from Efros et al. (2003), optical flow vectors were used to compute spatial motion descriptors. First, optical flow vectors were extracted between pixels of two consecutive frames. Then, spatial motion descriptors for frame pairs were encoded in three channels $F = (F_x, F_y, F_m)$ including x-motion, y-motion and magnitude. These descriptors were down-sampled to half resolution by skipping two pixels and normalized by mean frame subtraction. Finally, the Optical Flow descriptor of the video was the stacked spatial descriptors with x-motion, y-motion and magnitude channels.

Kanade-Lucas-Tomasi Trajectory model The Kanade-Lucas-Tomasi Trajectory (KLT) model is the representation of video stimuli to analyze the long-term motion pattern using trajectories. A long-term trajectory is the motion track of a local interest point in consecutive frames of the video and it models the movement kinematics of that point more precisely than the noisy optical flow vectors. Given a video stimulus, first 25 video clips each lasting 7 consecutive frames were extracted by sampling the middle frame at every 2 frames. Next, trajectories of local interest points sampled with a rate of 1000 were computed on each clip using Birchfield's implementation (Birchfield, 1998) of the KLT feature tracker (Shi and Tomasi, 1994). Then, each trajectory was represented in two channels including x-motion and y-motion. For a global representation of a clip, representations of trajectories were pooled by using a pyramid of spatial grid structure. Having a $m \times m$ spatial grid fitted as a layout on the middle (fourth) frame of the video clip, a single normalized trajectory feature was computed at each spatial bin location by averaging velocity values of all trajectories passing from a bin (this is a 2×6-dimensional feature vector for trajectories lasting 7 frames). Finally, the KLT descriptor of the video was the stacked features over a spatial pyramid in five scales, $m \in (1, 2, 4, 8, 16)$ and over all video clips.

Space-time Interest Points model Space-time Interest Points (STIPs) descriptor (Laptev, 2005) was introduced as a spatial-temporal feature on interest points extracted by 3D-Harris detector and it measures the variations on detected points in space and time domains with good performances for action recognition (Laptev et al., 2008). Extracting local space-time points in multiple-scales, two descriptors based on histogram representation were computed in the neighborhood of each point. The neighborhood was defined as a space-time volume V divided into a grid, m_x, m_y, m_t. For each grid location, a histogram of oriented gradients (*HoG*) (Dalal and Triggs, 2005) and a histogram of optical flow vectors (*HoF*) were computed to capture the local appearance and the local motion feature, respectively. Then, the normalized *HoG* histograms per bin were combined as the *STIP-hog* descriptor while the *HoF* histograms were combined as the *STIP-hof* descriptor. We used the software with default setting.[4]

Our action models were the classical BoF representations using STIPs descriptors on spatial-temporal interest points. For learning a

visual vocabulary of STIPs descriptors, we used HMDB (Kuehne et al., 2011) as training video dataset, since it consists of 51 human actions including upper body ones with object interactions. The original dataset consists of 7000 video clips, but we selected a random subset of 10 clips per action category for computing the vocabulary. First, the video clips of the HMDB subset were resized into frames of size 96×96 following our feature extraction setup. Then, 2000 cluster centers were computed as visual words over extracted STIPs descriptors using the k-means method. Similarly, STIPs descriptors were extracted from our stimuli set and each descriptor was assigned to a word of the visual vocabulary. Finally, ℓ_2 normalized histograms were computed for each video stimulus as the BoF representation. As action models, two BoF representations were obtained from both the *STIP-hog* and the *STIP-hof* descriptors.

Dense Trajectory model Dense Trajectory (DT) descriptor was first introduced as a video representation based on long-term trajectories sampled densely (Wang et al., 2011) and it still has comparable performance with features based on Convolutional Neural Network (CNN) for action recognition on complex video datasets (Tran et al., 2015).

For the DT descriptors, first trajectories were extracted as tracks of dense optical flow fields in multiple scales. To avoid the problem of trajectory drifting, the length of trajectories was fixed as lasting L frames. Next, four trajectory-aligned descriptors for each trajectory were computed to capture local information within a space-time volume V defined by $M \times M$ pixels and L frames. To capture the structure of the finer details, the volume V was divided into a grid, m_x, m_y, m_t. Similar to STIPs descriptors, the *HoG* and the *HoF* descriptors were extracted for each bin locations to capture local appearance and motion along the trajectory, respectively. Moreover, two motion-based descriptors, namely trajectory (*Trj*) descriptor and a motion boundary histogram (*MBH*) descriptor (Dalal et al., 2006) were computed. *Trj* descriptor was the motion pattern of the trajectory with L frames. Descriptor was the vector of normalized displacement values along the trajectory points. Similar to *HoF*, *MBH* descriptor was based on the optical flow vectors but on the derivatives of motion to suppress the fields with constant motion. Having flow vectors, the derivatives of x-motion and y-motion components were taken individually. The orientation information was aggregated into histograms as in the *HoF* descriptors for each motion channel. The *MBH* descriptor was the concatenated normalized histograms. To extract dense trajectory descriptors, we used the software with the default setting,[5] where $L = 15$, $M = 32$, $m_x = m_y = 2$, and $m_t = 3$.

Similar to STIPs action models, we developed models of BoF representations for each descriptor type of the DT model, namely *DT-trj*, *DT-hog*, *DT-hof* and *DT-mbh* descriptors. Using the same subset of HMDB dataset (Kuehne et al., 2011), we computed 2000 visual words for the corresponding type of DT descriptor and we extracted ℓ_2 normalized BoF histograms as the computer vision models of the video stimuli.

2.6.3. Attribute-based categorical models

For a high level representation of the video stimuli, we constructed the RDMs using attribute-based categorical models by modeling the action videos as attributes in a given category. There were 6 categorical models, and each categorical model included two or more attributes. The list of categorical models included: **Agent** (with attributes Robot, Android, Human), **Movement** (with attributes Biological, Nonbiological), **Appearance** (with attributes Biological, Nonbiological), **Intention** (with attributes Manipulative, Communicative, and Self-movement), **Target** (with attributes Human, Object), and **Action** (with attributes Drink, Grasp, Handwave, Talk, Nudge, Throw, Turn, Wipe). Although there may be other attribute-based models, we constructed these ones as they allowed us to easily model the high-level and semantic aspects of the video stimuli in an objective manner.

While modeling our video stimuli, each video was represented as a binary vector, and each component of the vector represented the membership of the video for the corresponding attribute. Thus, each video was labeled with **1** if it featured the attribute in a given categorical model and **0** otherwise (see Fig. 3 for the ground truth attribute labeling of the video stimuli). For each categorical model, a $N \times N$ RDM was constructed in which pair-wise dissimilarities between video representations were computed using the hamming distance.

2.6.4. Convolutional Neural Network (CNN) model

Mainly due to the advances of deep learning models in computer vision, the quality of recognition even had dramatic progress. More specifically, the deep models by Convolutional Neural Network (CNN) models had been shown to be more robust than the hand-crafted shallow models (e.g., HoG, HoF). More recent studies have demonstrated that deep representations were also possible for video interpretation (Karpathy et al., 2014; Tran et al., 2015; Carreira and Zisserman, 2017). Following these studies, we used CNN model as deep feature representations of the video stimuli.

We focused on the deep learning features using 3D Convolutional Neural Network model (C3D) (Tran et al., 2015), since it models appearance and motion simultaneously using $3 \times 3 \times 3$ convolutional kernel in contrary to the 2D neural network models developed for image domain. C3D was a deep model consisting of 8 convolutional, 5 max-pooling, and 2 fully connected layers. The pre-trained C3D model (using tensorflow implementation[6]) on the Sport-1M dataset (Karpathy et al., 2014) was used to extract deep activations on our video stimuli. We selected 5 pooling layers, 2 fully connected layers and the final softmax layer as our intermediate layers of video representation (see Fig. 4 for the network structure of the C3D model and sample frames with deep activation features). While the lower layers of the neural model can be thought as corresponding to the representation of low-level visual features, the higher layers can be though as high-level features (Zeiler and Fergus, 2014). For each layer of representation, a $N \times N$ RDM was constructed in which pair-wise dissimilarities between video representations were computed using the Spearman correlation.

3. Results

3.1. ROIs involved in action observation

Consistent with prior studies of action observation (Caspers et al., 2010), the visual action stimuli in the main experiment resulted in activation in the early visual cortex (EVC), extending to dorsal and ventral visual streams, and core nodes of the AON: pSTS, inferior parietal lobe, and ventral premotor cortex, which were identified by running the GLM and contrasting all the video stimuli with the fixation in the main experiment ($p < 0.001$ uncorrected) (see Fig. 5 and Table 1). Then, we extracted a 4 mm sphere ROI that covered the activation pattern in pSTS, inferior parietal, and ventral premotor cortex. Some subjects did not show activations in some ROIs of the AON with the common threshold we used for all subjects. More specifically, one subject's occipito-temporal activation was restricted with the MT cluster and did not extend to pSTS. In 4 subjects, although premotor level of the AON was activated, only the dorsal premotor part was active. To select a more homogeneous set of voxels and be consistent with the literature, we restricted our analysis with the subjects who had ventral premotor activation. We identified all three ROIs of AON (pSTS, inferior parietal, ventral premotor) in 18 subjects so the rest of the analysis included these subjects. The coordinates of the central voxels are provided in Table 2 for the ROIs that form the core nodes of the AON.

[4] https://www.di.ens.fr/laptev/actions/.
[5] http://thoth.inrialpes.fr/software.
[6] https://github.com/hx173149/C3D-tensorflow.

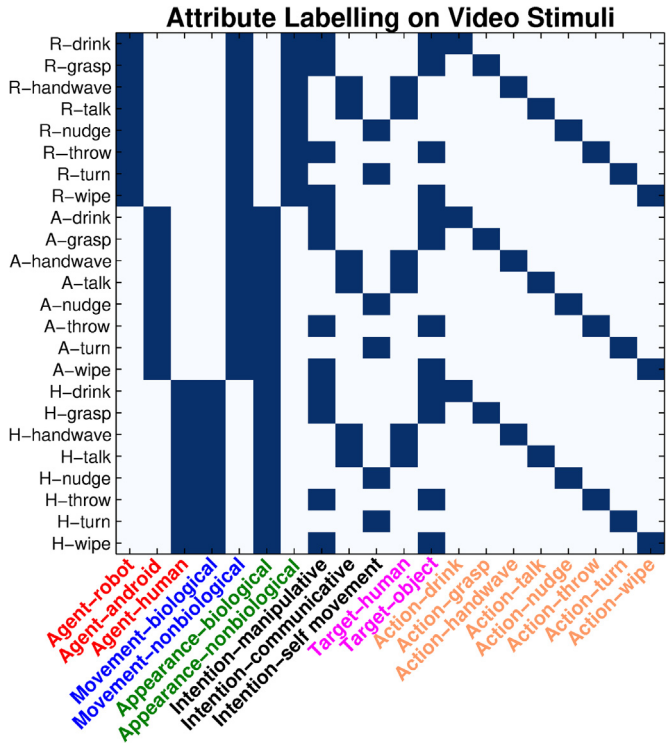

Fig. 3. Attribute-based categorical models as high level representation of video stimuli. Each video stimulus was represented in attributes of 6 categorical models including Agent, Movement, Appearance, Intention, Target, and Action.

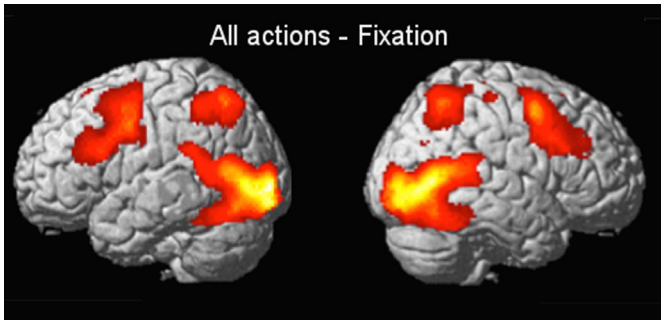

Fig. 5. Activation map for the observation of all actions compared to the fixation period at the group level ($p < 0.001$ uncorrected, cluster size = 5).

common voxels defined at the group level, which are not active at the individual subject level. That is, there may be variation between the group level activation and individual subject activation. Hence, the voxels that seem to be active at the group level but are not active at the individual subject level may introduce noise in the RSA. We indeed examined our data this way and saw that there were voxels that were found active at the group level but were not active in some subjects. We aimed to overcome this issue by defining the ROIs at the individual subject level. This approach is rather conservative so we ended up having ROIs that are defined with 4 mm radius sphere rather than larger ROIs that could be defined at the group level. On average, subjects were 96% accurate in the behavioral task.

3.2. RSA: brain RDMs

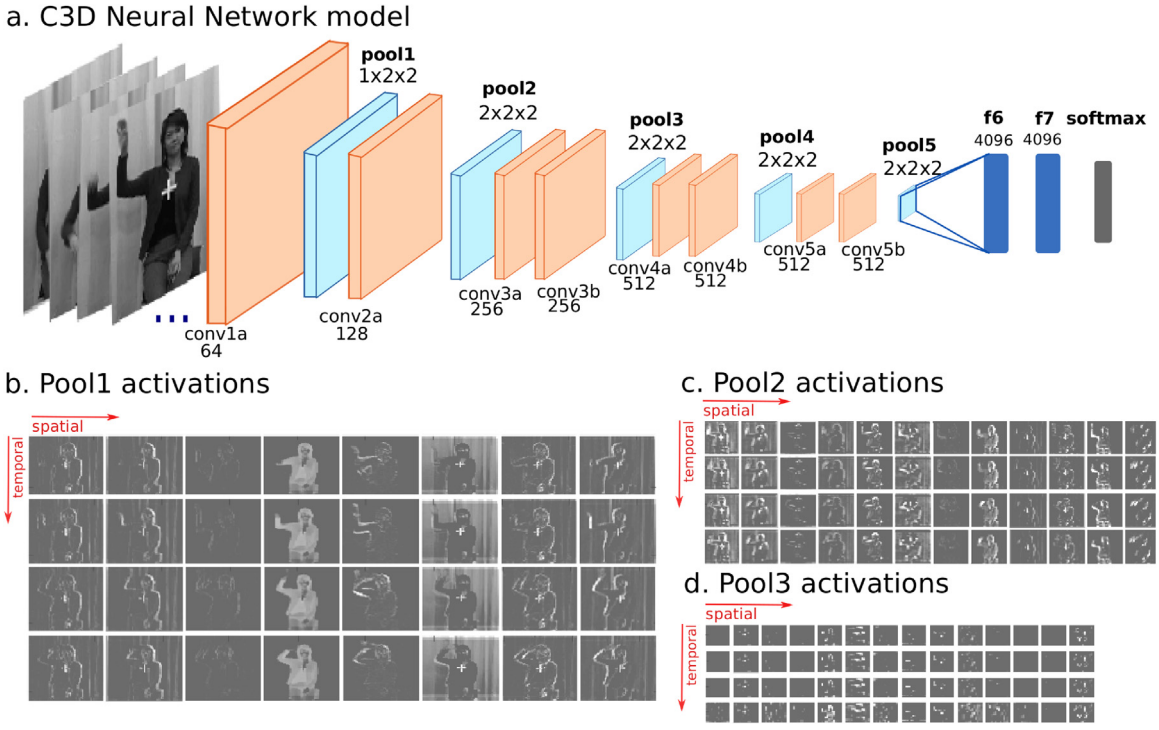

Fig. 4. 3D convolutional neural network model to represent video stimuli. (a) Structure of the 3D Convolutional Neural Network (C3D) model including convolutional, max-pooling, and fully connected layers, and (b) Sample activations from pool1, pool2 and pool3 intermediate layers on handwave video stimulus.

Please note that we defined the ROIs at the individual subject level for each level of the AON rather than using a common set of voxels across subjects. The primary motivation of doing this analysis was to overcome the disadvantage of the group-level defined ROIs. The disadvantage of this approach is that there may be voxels in the set of

The structure of the RDMs throughout the paper for the brain ROIs and models was shown in Fig. 6a. The brain RDMs for pSTS, parietal, and premotor nodes were shown in Fig. 6b-d. Each RDM shows the grand average of 18 subjects, and the two hemisphere representations were averaged for each ROI.

Table 1
MNI coordinates of the peak voxels of the brain regions involved in visual processing of actions based on the all actions – fixation contrast ($p < 0.001$ uncorrected, cluster size = 5) in the whole brain GLM analysis (see Fig. 5).

MNI coordinates				
x	y	z	Anatomical Name	Brodmann Area
−34	**−92**	**0**	Middle occipital gyrus (left)	BA 17
− 26	− 92	− 10	Inferior occipital gyrus (left)	BA 18
− 48	− 80	− 2	Middle occipital gyrus (left)	BA 19
48	**−74**	**−2**	Inferior temporal gyrus (right)	BA 19
40	− 84	− 8	Inferior occipital gyrus (right)	BA 19
22	− 94	− 6	Sub-gyral (right)	BA 18
42	**2**	**56**	Middle frontal gyrus (right)	BA 6
50	34	34	Middle frontal gyrus (right)	BA 9
46	10	30	Inferior frontal gyrus (right)	BA 9
−34	**−58**	**50**	Superior parietal lobule (left)	BA 39
38	**−56**	**52**	Inferior parietal lobule (right)	BA 40
32	− 68	28	Sub-gyral (right)	BA 39
−44	**0**	**56**	Middle frontal gyrus (left)	BA 6
− 42	− 2	38	Middle frontal gyrus ((left))	
− 60	6	32	Inferior frontal gyrus(left)	BA 6
−6	**12**	**50**	Medial frontal gyrus (left)	BA 6
28	**−6**	**−22**	Amygdala (right)	
−12	**26**	**60**	Superior frontal gyrus (left)	BA 6
8	**−22**	**68**	Medial frontal gyrus (right)	BA 6
38	**−26**	**58**	Precentral gyrus (right)	BA 4

models (see Fig. 8a): STIP-hof (*Kendalltau* = 0.028, $p = 0.0024$), DT-trj (*Kendalltau* = 0.022, $p = 0.011$), and DT-mbh (*Kendalltau* = 0.027, $p = 0.008$). Each of these three correlations exceeded the lower bound of the noise ceiling for pSTS, which was estimated as − 0.0074, suggesting that these three are candidate models that explain pSTS representation of observed actions. The correlation of pSTS with the other computer vision models was not significant (Gabor: *Kendalltau* = 0.0008, $p = 0.96$; Optic Flow: *Kendalltau* = 0.015, $p = 0.18$; KLT: *Kendalltau* = 0.013, $p = 0.17$; STIP-hog: *Kendalltau* = −0.007, $p = 0.37$; DT-hog: *Kendalltau* = −0.007, $p = 0.68$; DT-hof: *Kendalltau* = 0.012, $p = 0.19$). We further examined whether the three candidate models (STIP-hof, DT-trj, and DT-mbh) were significantly different from each other to find out whether one of these models was a better model than the others for pSTS. We employed pairwise *t*-tests between the three model correlations with pSTS (corrected for multiple comparisons $p < 0.016$). The correlations of these models with pSTS were not significantly different from each other (STIP-hof vs. DT-trj: $p = 0.59$; STIP-hof vs. DT-mbh: $p = 0.86$; DT-trj vs. DT-mbh: $p = 0.49$).

The parietal node was significantly correlated with only one computer vision model (see Fig. 8b): DT-hog (*Kendalltau* = −0.026, $p = 0.011$). However, this correlation did not exceed the lower bound of the noise ceiling for the parietal node, which was estimated as 0.0052. The correlation of the parietal node with the rest of the computer vision

Table 2
Average MNI coordinates of the central voxels of the extracted ROIs in the core Action Observation Network. The values in parenthesis indicate the standard deviation.

	CORE NODES OF ACTION OBSERVATION NETWORK									
	Left					Right				
	x	y	z	Sphere Size (radius in mm)	Anatomical Name	x	y	z	Sphere Size (radius in mm)	Anatomical Name
pSTS	− 50	− 53	7	4	Superior	53	−46	8	4	Superior
	(5	9	4)		Temporal Gyrus	(5	7	4)		Temporal Gyrus
Parietal	− 31	− 55	47	4	Inferior	35	−54	49	4	Superior
	(4	5	7)		Parietal Lobule	(6	5	6)		Parietal Lobule
Premotor	− 44	7	28	4	Inferior	46	9	27	4	Inferior
	(4	7	5)		Frontal Gyrus	(5	9	4)		Frontal Gyrus

3.3. RSA: computer vision models and brain relation

We linked the RDM of the computer vision models (see Fig. 7) with that of the core nodes of the AON, pSTS, inferior parietal lobe, and ventral premotor cortex (see Fig. 6) via Kendall tau correlation (corrected for multiple comparisons with FDR, $p < 0.016$ as explained in Section 2.6). The correlation results are shown in Fig. 8.

pSTS was significantly correlated with three computer vision models was insignificant (Gabor: *Kendalltau* = −0.002, $p = 0.49$; Optic Flow: *Kendalltau* = 0.004, $p = 0.49$; KLT: *Kendalltau* = 0.003, $p = 0.49$; STIP-hog: *Kendalltau* = −0.02, $p = 0.054$; STIP-hof: *Kendalltau* = 0.005, $p = 0.49$; DT-trj: *Kendalltau* = 0.005, $p = 0.49$; DT-hof: *Kendalltau* = −0.009 $p = 0.49$; DT-mbh: *Kendalltau* = −0.006, $p = 0.73$). This suggests that none of the computer vision models is a candidate model to explain the parietal cortex representation of observed actions.

Ventral premotor cortex was not significantly correlated with any of

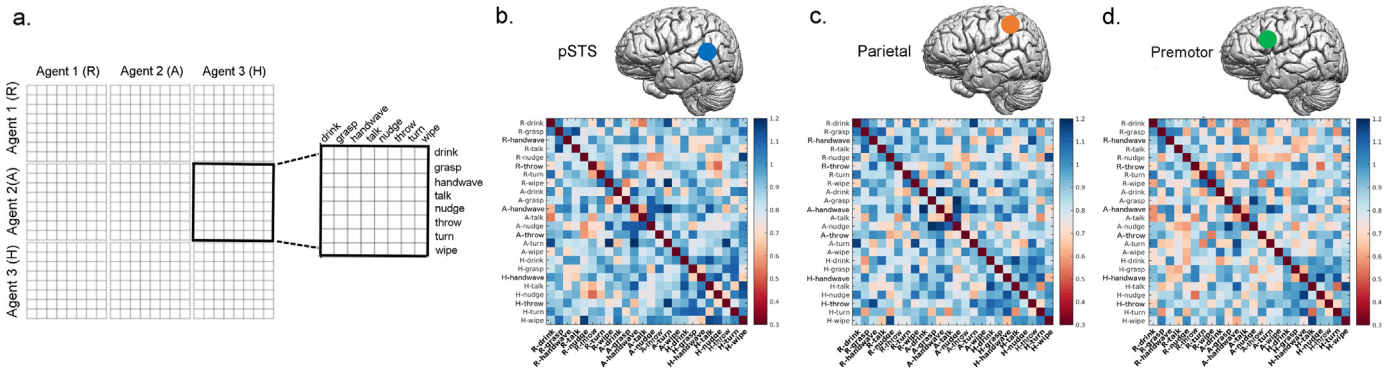

Fig. 6. RDMs for the three nodes of the AON. (a) The structure of the RDMs in (b-d), (b) RDM for the pSTS, (c) RDM for the parietal node, (d) RDM for the premotor node of the AON.

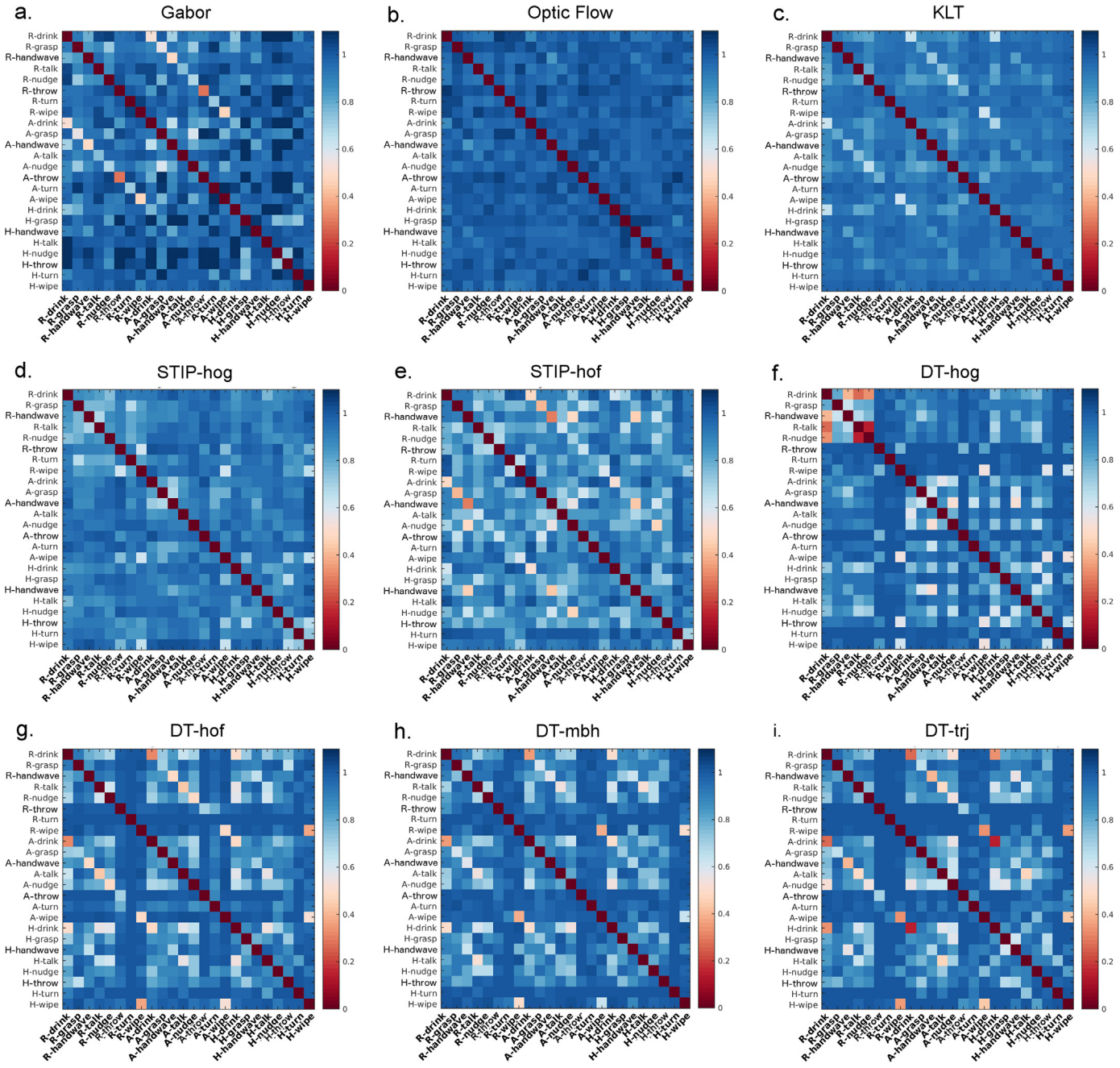

Fig. 7. RDMs for the computer vision models. (a) Gabor, (b) Optical Flow, (c) KLT, (d) STIP-hog, (e) STIP-hof, (f) DT-hog, (g) DT-hof, (h) DT-mbh, (i) DT-trj.

the computer vision models (see Fig. 8c): (Gabor: *Kendalltau* = 0.002, $p = 0.92$; Optic Flow: *Kendalltau* = 0.006, $p = 0.80$; KLT: *Kendalltau* = 0.01, $p = 0.80$; STIP-hog: *Kendalltau* = −0.01, $p = 0.11$; STIP-hof: *Kendalltau* = 0.006, $p = 0.79$; DT-trj: *Kendalltau* = 0.02, $p = 0.33$; DT-hog: *Kendalltau* = −0.01, $p = 0.33$; DT-hof: *Kendalltau* = 0.008, $p = 0.80$; DT-mbh: *Kendalltau* = 0.008, $p = 0.80$). This suggests that none of the computer vision models was a candidate model to explain the ventral premotor cortex representation of observed actions.

3.4. RSA: attribute-based categorical models and brain relation

We linked the RDMs of the categorical models (see Fig. 9) with that of the core nodes of the AON, pSTS, inferior parietal, and ventral premotor cortex (see Fig. 6) via Kendall tau correlation (corrected for multiple comparisons with FDR, $p < 0.016$ as explained in Section 2.6). The correlation results are shown in Fig. 10.

pSTS was significantly correlated with three categorical models (see Fig. 10a): Agent (*Kendalltau* = −0.028, $p = 0.0005$), Appearance (*Kendalltau* = −0.036, $p = 0.0004$), and Action (*Kendalltau* = 0.023, $p = 0.003$). However, only the correlation with the Action categorical model exceeded the lower bound of the noise ceiling estimated for pSTS (− 0.0074) ruling out the Agent and Appearance categorical models as candidate models for pSTS representation of observed actions. The correlation of pSTS with the rest of the categorical models was not significant (Target: *Kendalltau* = 0.007, $p = 0.24$; Movement: *Kendalltau* = −0.003, $p = 0.38$; Intention: *Kendalltau* = 0.009, $p = 0.24$).

The parietal node of the AON significantly correlated with all of the categorical models (see Fig. 10b) (Target: *Kendalltau* = 0.021, $p = 0.004$; Agent: *Kendalltau* = −0.035, $p = 0.000001$; Appearance: *Kendalltau* = −0.014, $p = 0.0003$; Intention: *Kendalltau* = 0.02, $p = 0.0022$; Action: *Kendalltau* = 0.029, $p = 0.00017$) except the

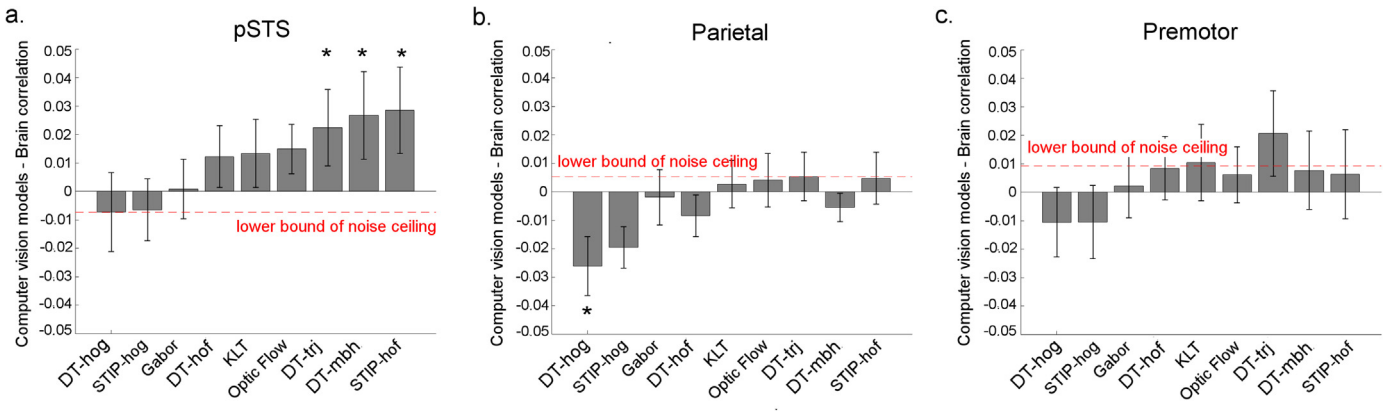

Fig. 8. The correlations between the computer vision models and the three core nodes of the AON. (a) Correlations with the pSTS, (b) Correlations with the parietal node, (c) Correlations with the premotor node. Asterisk (*) indicates significant correlations ($p < 0.05$ corrected with FDR). The red dotted line in each panel indicates the estimated lower bound of the noise ceiling for each region. Any model that has a significant correlation with a given region and that exceeds the lower bound of the noise ceiling is considered to be a candidate model representation of the respective ROI.

Movement model (*Kendalltau* = −0.019, $p = 0.03$). However, among these models, the correlation with only Target, Intention, and Action models exceeded the lower bound of the noise ceiling estimated for the parietal node (0.0052), suggesting that these three are the candidate models to explain the parietal representation of observed actions. In order to find out whether one of these models was a better model than the other, we employed pair-wise *t*-tests between the three model correlations with the parietal node (corrected for multiple comparisons, $p < 0.016$). The models correlations with the parietal node did not significantly differ from each other (Target-Intention: $p = 0.90$; Target-Action: $p = 0.41$; Intention-Action: $p = 0.55$).

To further investigate whether a combination of one or more of these models better correlated with the parietal node, we defined 4 new categorical models by combining the binary representation of the earlier categorical models and constructing their RDMs: Target-Intention, Target-Action, Intention-Action, and Target-Intention-Action. We then correlated the RDM of each model with that of the parietal node. We found that all of these models significantly correlated with the parietal node (see Fig. 11) (Target-Intention: *Kendalltau* = 0.021, $p = 0.004$; Target-Action: *Kendalltau* = 0.023, $p = 0.0009$; Intention-Action: *Kendalltau* = 0.024, $p = 0.0008$; Target-Intention-Action: *Kendalltau* = 0.024, $p = 0.0009$) but they did not significantly differ from each other

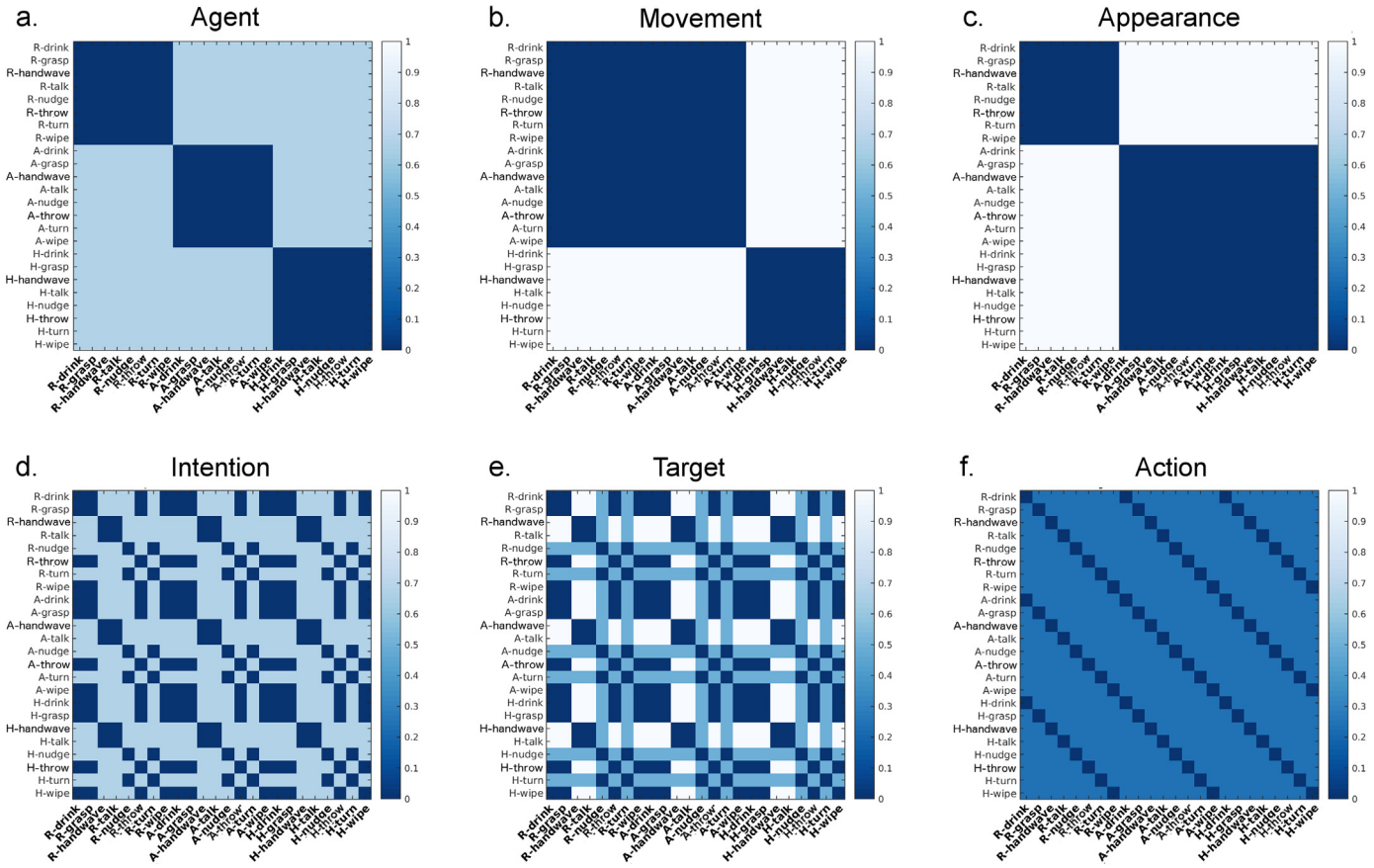

Fig. 9. RDMs for the categorical models. (a) Agent, (b) Movement, (c) Appearance, (d) Intention, (e) Target, (f) Action.

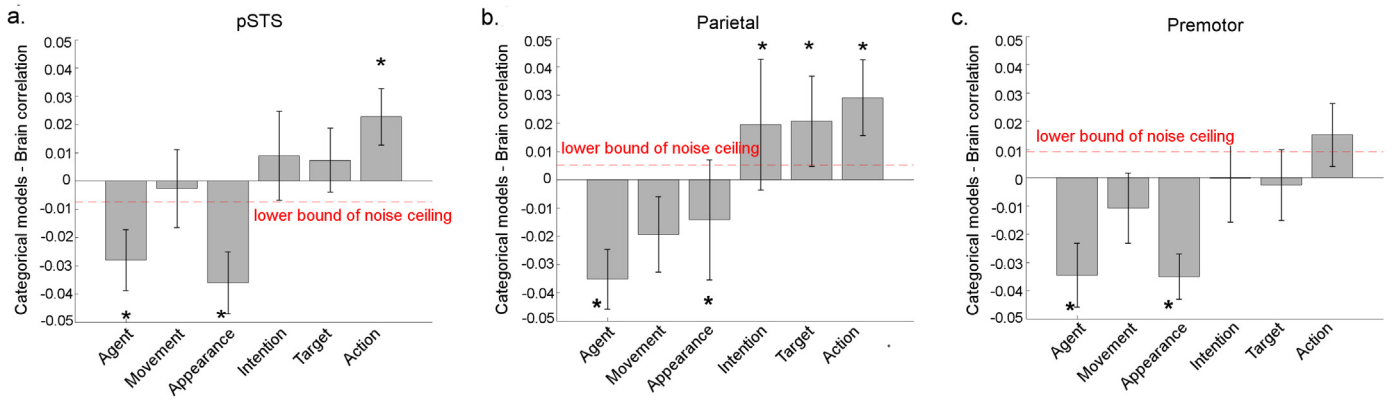

Fig. 10. The correlations between the categorical models and the three core nodes of the AON. (a) Correlations with the pSTS, (b) Correlations with the parietal node, (c) Correlations with the premotor node. Asterisk (*) indicates significant correlations ($p < 0.05$ corrected with FDR). The red dotted line in each panel indicates the estimated lower bound of the noise ceiling for each region. Any model that has a significant correlation with a given region and that exceeds the lower bound of the noise ceiling is considered to be a candidate model representation of the respective ROI.

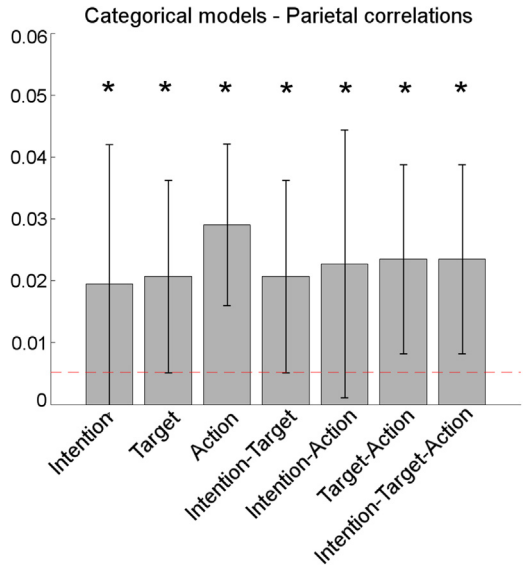

Fig. 11. The correlations between the basic categorical models that significantly correlate with the parietal node of the AON, and their combinations. Asterisk (*) indicates significant correlations ($p < 0.05$ corrected with FDR). The red dotted line indicates the estimated lower bound of the noise ceiling.

($p > 0.05/6 = 0.008$, corrected for the 6 comparisons at $p = 0.05$). However, all these correlations exceeded the lower bound of the noise ceiling estimated for the parietal node, suggesting that these models are candidate models to explain the parietal cortex representation of the observed actions.

The ventral premotor cortex was significantly correlated with only the Agent model (see Fig. 10c) (*Kendalltau* = −0.035, $p = 0.0000025$) and Appearance model (*Kendalltau* = −0.035, $p = 0.00012$). However, the correlation with these models did not exceed the lower bound of the noise ceiling estimated for the ventral premotor cortex (0.0092). The correlation of the ventral premotor cortex with the rest of the categorical models was not significant (Target: *Kendalltau* = −0.003, $p = 0.98$; Movement: *Kendalltau* = −0.01, $p = 0.09$; Intention: *Kendalltau* = −0.00001, $p = 0.71$; Action: *Kendalltau* = 0.015, $p = 0.05$), suggesting that none of the categorical models is a candidate model to explain the ventral premotor cortex representation of the observed actions.

3.5. RSA: CNN model and brain relation

We linked the RDM of the selected layers of the CNN model for action recognition (see Fig. 12a) with that of the core nodes of the AON, pSTS, inferior parietal lobe, and ventral premotor cortex (see Fig. 6) via Kendall tau correlation (corrected for multiple comparisons with FDR, $p < 0.016$ as explained in Section 2.6). The correlation results are shown in Fig. 12b.

pSTS was significantly negatively correlated with all layers of the CNN except Pool2 (*Kendalltau* = −0.01, $p = 0.07$) and f7 (*Kendalltau* = −0.013, $p = 0.18$) (Pool1: *Kendalltau* = −0.019, $p = 0.006$; Pool3: *Kendalltau* = −0.017, $p = 0.009$; Pool4: *Kendalltau* = −0.02, $p = 0.005$; Pool5: *Kendalltau* = −0.028, $p = 0.0004$; f6: *Kendalltau* = −0.027, $p = 0.0004$; Softmax: *Kendalltau* = −0.029, $p = 0.0004$). However, none of the layer correlations exceeded the noise ceiling estimated for pSTS (− 0.0074), suggesting that none of the layers of the CNN is a good model for the pSTS representation of observed actions.

The parietal node was significantly negatively correlated with all layers of the CNN except f7 (*Kendalltau* = −0.009, $p = 0.22$) (Pool1: *Kendalltau* = −0.021, $p = 0.00001$; Pool2: *Kendalltau* = −0.019, $p = 0.00006$; Pool3: *Kendalltau* = −0.02, $p = 0.00004$; Pool4: *Kendalltau* = −0.019, $p = 0.00008$; Pool5: *Kendalltau* = −0.019, $p = 0.00006$; f6: *Kendalltau* = −0.01, $p = 0.0008$; Softmax: *Kendalltau* = −0.01, $p = 0.012$). However, none of the layer correlations exceeded the noise ceiling estimated for the parietal node (0.0052), suggesting that none of the layers of the CNN is a good model for the parietal node representation of observed actions.

The ventral premotor node was significantly negatively correlated with all layers of the CNN (Pool1: *Kendalltau* = −0.03, $p = 0.00001$; Pool2: *Kendalltau* = −0.024, $p = 0.00006$; Pool3: *Kendalltau* = −0.027, $p = 0.00006$; Pool4: *Kendalltau* = −0.025, $p = 0.0002$; Pool5: *Kendalltau* = −0.031, $p = 0.00005$; f6: *Kendalltau* = −0.018, $p = 0.0008$; f7: *Kendalltau* = −0.021, $p = 0.003$; Softmax: *Kendalltau* = −0.027, $p = 0.0004$). However, none of the layer correlations exceeded the noise ceiling estimated for the premotor node (0.0092), suggesting that none of the layers of the CNN is a good model for the premotor node representation of observed actions.

So, overall we found that layers of C3D, a particular CNN implementation for action recognition are not plausible models for the representation of the AON.

4. Discussion

In the present study, we investigated what aspects of visually processed actions are represented in the human Action Observation

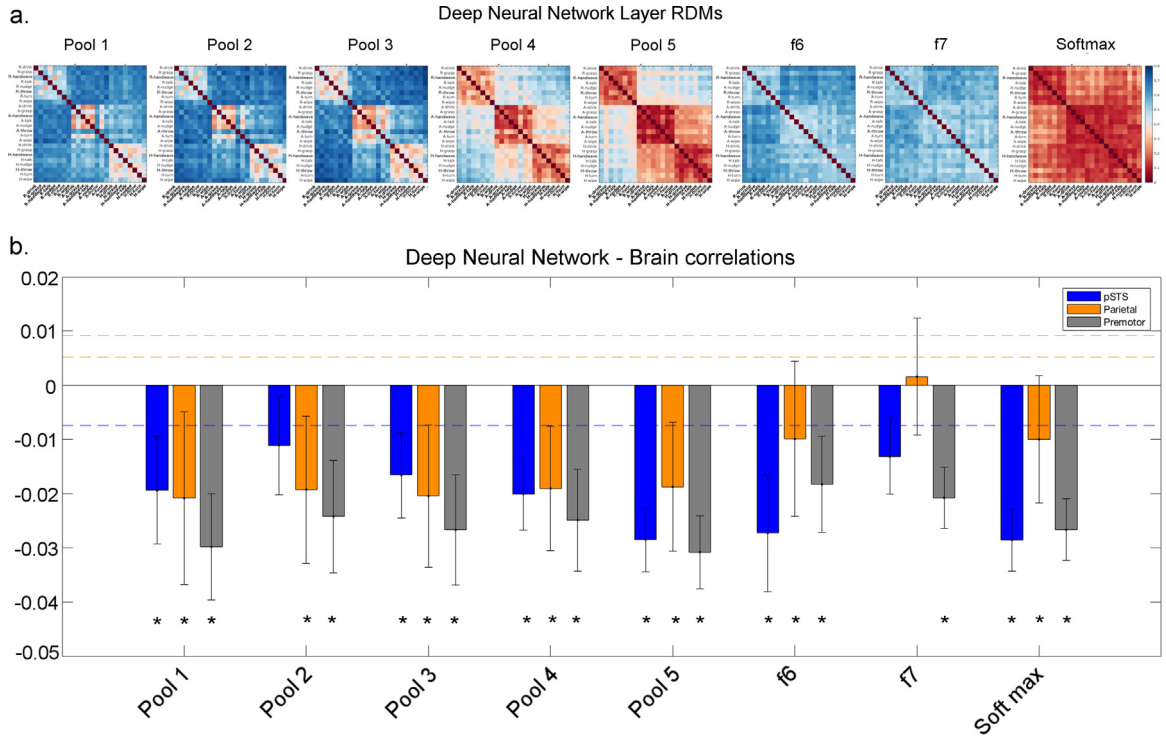

Fig. 12. RDMs for the 3D convolutional neural network (C3D) model and the correlations between network layers and the nodes of the AON. (a) The RDMs of the layers of the deep neural network (b) The correlation of the layers of the C3D with the three nodes of the AON. Asterisk (*) indicates significant correlations ($p < 0.05$ corrected with FDR). The dotted lines indicate the estimated lower bound of the noise ceiling for each region.

Network. To this end, we combined computer vision models with categorical models of the action stimuli, and linked them to brain responses using representational similarity analysis. Computer vision models characterized the visual form and motion information in the action stimuli with varying degrees of complexity, whereas categorical models defined the high-level visual and semantic aspects of actions. Our results show that different aspects of actions are represented in distinct nodes of the Action Observation Network revealing their representational differences.

4.1. pSTS

pSTS is the most suitable candidate region that could link the Action Observation Network to the early visual cortex, given its anatomical proximity to the early visual areas that process form and motion information. Thus, we hypothesized that pSTS would correlate well with computer vision models that characterize the form and motion aspects of actions rather than the categorical models that characterize the higher-aspects of actions. We found evidence that this is indeed the case.

We were further interested in identifying what kind of a vision model can best represent the pSTS responses. Among the computer vision models, pSTS correlated well only with models that characterize form and motion aspects at a high-degree of complexity (STIP and DT models) compared with the ones that have lower-complexity (KLT) or characterize only form (Gabor) or only motion information (Optical Flow). A closer look at the specific implementation of those well-correlated models suggest that pSTS does form processing by means of identifying the local interest points (e.g. body parts) in the observed action followed by motion processing within those interest points by means of tracking their trajectories. These results are consistent with the proposals that pSTS is the node where form and motion information are integrated during perception of body movements (Vaina et al., 2001; Giese and Poggio, 2003; Thompson and Baccus, 2012; Tan et al., 2013; Theusner et al., 2014).

Once we established that pSTS correlated well with the computer vision models, we further investigated whether it would correlate well with a deep neural network trained for action recognition (C3D) given their increasing application in neuroscience, and if so whether its correlation would be better than that of the parietal and premotor cortex. We found no evidence that pSTS performs similar to a deep neural network since the correlations with neither layer exceeded the lower bound of the noise ceiling. Furthermore, pSTS did not differ from parietal and premotor cortex in terms of the correlations with the neural network layers. We also tested two other deep neural networks, one image-based (VGG19 (Simonyan and Zisserman, 2014)) and one video-based (I3D (Carreira and Zisserman, 2017)), and they behaved similarly to C3D (therefore we did not report their results here).

Among the categorical models, Action model that defines the action type was the only one that represent the pSTS response patterns. These results are consistent with and extends recent empirical evidence that action type can be decoded from the occipito-temporal node of the Action Observation Network (Tucciarelli et al., 2015; Hafri et al., 2017). Furthermore, they indicate that the most complex computer vision models that correlated well with pSTS (STIP and DT as indicated above) may actually code the action type by means of integrating form and motion at a high-level.

Taken together, our results suggest that a deep neural network cannot account for the representations of pSTS as the visual node of the Action Observation Network. Rather, hand-crafted visual features such as STIP and DT that characterize the integration of form and motion information better represent pSTS response patterns. However, it is possible that lower regions in the action-relevant visual hierarchy (e.g. EBA) behave similar to a deep neural network, which can be tested in future studies.

4.2. Parietal node

Our results show that parietal node of the Action Observation Network correlated well with categorical models and not with

computer vision models that define form and motion aspects of the action stimuli. More importantly, among the categorical models the ones that are based on purely visual properties such as movement kinematics, appearance, or type of the agent were found to be not good models for parietal cortex. Instead, the models that define either some high-level aspects such as the target or type of the action or more abstract and semantic aspects such as the intention of the action were found to be good models.

The results for the Target model suggest that parietal cortex is not only sensitive to the physical interaction of an effector (e.g. hand) with an object, as has been shown earlier for simple manipulative actions (Fleischer et al., 2013; Fabbri et al., 2016), but codes for any interaction between the effector and the target, regardless of it being an object or a biological agent (human) who is not in physical contact with the effector. These results are consistent with and extends recent evidence that shows that parietal cortex generalizes across the object category the hand is interacting with Wurm and Lingnau (2015).

On the other hand, the results for the Action type and Intention models show evidence that parietal cortex is functionally organized to code for different actions both at the exemplar level (individual actions) and abstract class level (grouping of actions based on their intentions such as manipulative or communicative). These results are consistent with a number of fMRI studies that show that (1) parietal cortex hosts separate anatomical regions for different action classes in action observation (Abdollahi et al., 2013; Ferri et al., 2015; Corbo and Orban, 2017), (2) abstract aspects of actions, and even intentions can be decoded from parietal cortex (Gallivan et al., 2011; Wurm and Lingnau, 2015; Hafri et al., 2017; Chen et al., 2017). Furthermore, they also indicate that parietal cortex has an intention-based functional organization for observed actions as for planned actions (Anderson and Buneo, 2002).

Taken together, our results suggest that parietal level of the Action Observation Network builds on the visual representations of the pSTS and codes high-level visual and abstract aspects of actions. In this way, it possibly contributes to the construction of the meaning of actions consistent with neuropsychological literature, which reports deficits in action comprehension skills as opposed to action perception skills in patients with parietal damage (Binder et al., 2017).

4.3. Premotor node

One of the well-known properties of the Action Observation Network is that its premotor node has an effector-specific functional organization (Fujii et al., 2008; Jastorff et al., 2010; Di Dio et al., 2013; Fabbri et al., 2016). However, given its visual responsiveness to actions, we investigated whether it also codes visual or higher-level aspects of actions beyond effector information. We did not find evidence of coding neither the purely visual aspects such as form and motion nor the high-level visual or abstract aspects including movement kinematics, appearance, agent, target, action, or intention. These results seem to be not consistent with the results of some studies (Johnson-Frey et al., 2003; Gallivan et al., 2011) who found that premotor cortex was sensitive to the goal or intentions of the actions. This discrepancy may be due to the differences between the action stimuli used in the two studies (they used only manipulative actions whereas we had a more variety including communicative actions) and the corresponding areas in premotor cortex. Nevertheless, our results clearly distinguish premotor cortex from the occipito-temporal and parietal levels, which correlated well with computer vision models and categorical models, respectively. Furthermore, they call for search for other types of models for premotor cortex in future research. One such model is an effector model, which could not be tested in the current study since the action stimuli did not vary in terms of effectors. Another possibility is a more complex model inspired from motor control which codes some basic motor primitives.

4.4. Conclusion

In sum, the present study shows that different aspects of actions are represented in distinct nodes of the Action Observation Network. Importantly, it provides strong evidence that Action Observation Network is a hierarchical system in which increasing levels of the cortex code increasingly complex aspects of actions, as proposed by theoretical accounts (Kilner et al., 2007) and empirical findings (Grafton and Hamilton, 2007). While occipito-temporal level (pSTS) performs the visual analysis of actions by integrating form and motion information passed through the visual cortex, parietal cortex as the next level of the hierarchy transforms the visual information into a more abstract and semantic information. This information is possibly transformed into a motor code in premotor cortex to realize how the action is physically performed.

CRediT authorship contribution statement

Burcu A. Urgen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review & editing. **Selen Pehlivan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review & editing. **Ayse P. Saygin:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgments

This research was supported by NSF (CAREER BCS1151805), DARPA, Kavli Institute for Brain and Mind, and Qualcomm Institute (Calit2). We thank Hiroshi Ishiguro and Intelligent Robotics Laboratory at Osaka University for the preparation of the stimuli, and Edward Nguyen for assistance in data collection.

References

Abdollahi, R.O., Jastorff, J., Orban, G.A., 2013. Common and segregated processing of observed actions in human spl. Cereb. Cortex 23, 2734–2753.
Anderson, R.A., Buneo, C.A., 2002. Intentional maps in posterior parietal cortex. Annu. Rev. Neurosci. 25, 89–220.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodol.) 289–300.
Binder, E., Dovern, A., Hesse, M.D., Ebke, M., Karbe, H., Saliger, J., Fink, G.R., Weiss, P.H., 2017. Lesion evidence for a human mirror neuron system. Cortex 90, 125–137.
Birchfield, S., 1998. An implementation of the kanade-lucas-tomasi feature tracker.
Blake, R., Shiffrar, M., 2007. Perception of human motion. Annu. Rev. Psychol. 58, 47–73.
Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'17) 2017. pp. 4724–4733.
Caspers, S., Zilles, K., Laird, A.R., Eickhoff, S.B., 2010. Ale meta-analysis of action observation and imitation in the human brain. Neuroimage 50, 1148–1167.
Chen, Q., Garcea, F.E., Jacobs, R.A., Mahon, B.Z., 2017. Abstract representations of object-directed action in the left inferior parietal lobule. Cereb. Cortex 28, 2162–2174.
Clarke, A., Tyler, L.K., 2014. Object-specific semantic coding in human perirhinal cortex. J. Neurosci. 34, 4766–4775.
Corbo, D., Orban, G.A., 2017. Observing others speak or sing activates spt and neighboring parietal cortex. J. Cogn. Neurosci. 29, 1002–1021.
Cross, E.S., Kraemer, D.J., Hamilton, A.F.d.C., Kelley, W.M., Grafton, S.T., 2008. Sensitivity of the action observation network to physical and observational learning. Cereb. Cortex 19, 315–326.
Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance. In: Proceedings of European Conference on Computer Vision. pp. 428–441.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893.
Di Dio, C., Di Cesare, G., Higuchi, S., Roberts, N., Vogt, S., Rizzolatti, G., 2013. The neural correlates of velocity processing during the observation of a biological effector in the parietal and premotor cortex. Neuroimage 64, 425–436.
Efros, A.A., Berg, A.C., Mori, G., Malik, J., 2003. Recognizing action at a distance. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 726–733.
Fabbri, S., Stubbs, K.M., Cusack, R., Culham, J.C., 2016. Disentangling representations of object and grasp properties in the human brain. J. Neurosci. 36, 7648–7662.
Fei-Fei, L., & Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories In: Proceedings of IEEE Computer Society Conference on Computer Vision

and Pattern Recognition (CVPR'05). pp. 524–531.
Ferri, S., Rizzolatti, G., Orban, G.A., 2015. The organization of the posterior parietal cortex devoted to upper limb actions: an fmri study. Hum. Brain Mapp. 36, 3845–3866.
Fleischer, F., Caggiano, V., Thier, P., Giese, M.A., 2013. Physiologically inspired model for the visual recognition of transitive hand actions. J. Neurosci. 33, 6563–6580.
Fujii, N., Hihara, S., Iriki, A., 2008. Social cognition in premotor and parietal cortex. Soc. Neurosci. 3, 250–260.
Gallivan, J.P., McLean, D.A., Valyear, K.F., Pettypiece, C.E., Culham, J.C., 2011. Decoding action intentions from preparatory brain activity in human parieto-frontal networks. J. Neurosci. 31, 9599–9610.
Giese, M.A., Poggio, T., 2003. Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. 4, 179–192.
Grafton, S.T., Hamilton, A.F.d.C., 2007. Evidence for a distributed hierarchy of action representation in the brain. Hum. Mov. Sci. 26, 590–616.
Hafri, A., Trueswell, J.C., Epstein, R.A., 2017. Neural representations of observed actions generalize across static and dynamic visual input. J. Neurosci. 37, 3056–3071.
Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: a survey. Image Vision. Comput. 60, 4–21.
Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., Orban, G.A., 2010. Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. J. Neurophysiol. 104, 128–140.
Johnson-Frey, S.H., Maloof, F.R., Newman-Norlund, R., Farrer, C., Inati, S., Grafton, S.T., 2003. Actions or hand-object interactions? Human inferior frontal cortex and action observation. Neuron 39, 1053–1058.
Jozwik, K.M., Kriegeskorte, N., Mur, M., 2016. Visual features as stepping stones toward semantics: explaining object similarity in it and perception with non-negative least squares. Neuropsychologia 83, 201–226.
Jozwik, K.M., Kriegeskorte, N., Storrs, K.R., Mur, M., 2017. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. Front. Psychol. 8, 1726.
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L., 2014. Large-scale video classi cation with convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2014 1725 1732.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput. Biol. 10, e1003915.
Kilner, J.M., Friston, K.J., Frith, C.D., 2007. Predictive coding: an account of the mirror neuron system. Cogn. Process. 8, 159–166.
Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T., 2011. HMDB: alarge video database for human motion recognition. In: Proceedings of International Conference on Computer Vision.
Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vision. 64, 107–123.
Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies. CVPR 1–8.
Mishkin, M., Ungerleider, L.G., 1982. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behav. Brain Res. 6, 57–77.
Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.A., 2011. Action observation circuits in the macaque monkey cortex. J. Neurosci. 31, 3743–3756.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553.
Olshausen, B.A., et al., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609.
Ritchie, J.B., Bracci, S., de Beeck, H.O., 2017. Avoiding illusory effects in representational similarity analysis: what (not) to do with the diagonal. Neuroimage 148, 197–200.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu. Rev. Neurosci. 27, 169–192.
Saygin, A.P., Cicekli, I., 2002. Pragmatics in human-computer conversations. J. Pragmat. 34, 227–258.
Saygin, A.P., Chaminade, T., Ishiguro, H., Driver, J., Frith, C., 2012. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Social. Cogn. Affect. Neurosci. 7, 413–422.
Shi, J., & Tomasi, C., 1994. Good features to track. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 593–600.
Simonyan, K., & Zisserman, A., 2014. Very deep convolutional networks for770 large-scale image recognition. CoRR, abs/1409.1556.
Tan, C., Singer, J.M., Serre, T., Sheinberg, D., Poggio, T., 2013. Neural representation of action sequences: how far can a simple snippet-matching model take us? Adv. Neural Inform. Process. Syst. 593–601.
Theusner, S., de Lussanet, M., Lappe, M., 2014. Action recognition by motion detection in posture space. J. Neurosci. 34, 909–921.
Thompson, J.C., Baccus, W., 2012. Form and motion make independent contributions to the response to biological motion in occipitotemporal cortex. Neuroimage 59, 625–634.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M., 2015. Learning spatio-temporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497.
Tucciarelli, R., Turella, L., Oosterhof, N.N., Weisz, N., Lingnau, A., 2015. Meg multi-variate analysis reveals early abstract action representations in the lateral occipito-temporal cortex. J. Neurosci. 35, 16034–16045.
Urgen, B.A., Plank, M., Ishiguro, H., Poizner, H., Saygin, A.P., 2013. Eeg theta and mu oscillations during perception of human and robot actions. Front. Neurorobotics 7, 10–3389.
Vaina, L.M., Solomon, J., Chowdhury, S., Sinha, P., Belliveau, J.W., 2001. Functional neuroanatomy of biological motion perception in humans. Proc. Natl. Acad. Sci. USA 98, 11656–11661.
Wang, H., Klaser, A., Schmid, C., & Liu, C.-L., 2011. Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Wurm, M.F., Lingnau, A., 2015. Decoding actions at different levels of abstraction. J. Neurosci. 35, 7727–7735.
Zeiler, M.D., Fergus R., 2014. Visualizing and understanding convolutional networks. In: Proceedings of European Conference on Computer Vision. pp. 818–833.