Social-affective features drive human representations of observed actions

Diana C. Dima[1*], Tyler Tomita[2], Christopher Honey[2], Leyla Isik[1]

[1]Dept. of Cognitive Science, Johns Hopkins University, United States

[2]Dept. of Psychological and Brain Sciences, Johns Hopkins University, United States

Abstract

Humans observe actions performed by others in many different visual and social settings. What features do we extract and attend when we view such complex scenes, and how are they processed in the brain? To answer these questions, we curated two large-scale sets of naturalistic videos of everyday actions and estimated their perceived similarity in two behavioral experiments. We normed and quantified a large range of visual, action-related and social-affective features across the stimulus sets. Using a cross-validated variance partitioning analysis, we found that social-affective features predicted similarity judgments better than, and independently of, visual and action features in both behavioral experiments. Next, we conducted an electroencephalography (EEG) experiment, which revealed a sustained correlation between neural responses to videos and their behavioral similarity. Visual, action, and social-affective features predicted neural patterns at early, intermediate and late stages respectively during this behaviorally relevant time window. Together, these findings show that social-affective features are important for perceiving naturalistic actions, and are extracted at the final stage of a temporal gradient in the brain.

* Correspondence should be addressed to: ddima@jhu.edu/diana.c.dima@gmail.com

Introduction

In daily life, we rely on our ability to recognize a range of actions performed by others in a variety of different contexts. Our perception of others' actions is both efficient and flexible, enabling us to rapidly understand new actions no matter where they occur or who is performing them. This understanding plays a part in complex social computations about the mental states and intentions of others (Jamali et al., 2021; Spunt et al., 2011; Thornton et al., 2019; Thornton and Tamir, 2021; Weaverdyck et al., 2021). Visual action recognition also interacts cross-modally with language-based action understanding (Bedny and Caramazza, 2011; Humphreys et al., 2013). However, there are two important gaps in our understanding of action perception in realistic settings. First, we still don't know which features of the visual world underlie our representations of observed actions. Second, we do not know how different types of action-relevant features, ranging from visual to social, are processed in the brain, and especially how they unfold over time. Answering these questions can shed light on the computational mechanisms that support action perception. For example, are different semantic and social features extracted in parallel or sequentially?

Relatively few studies have investigated the temporal dynamics of neural responses to actions. During action observation, a distributed network of brain areas extracts action-related features ranging from visual to abstract, with viewpoint-invariant responses emerging as early as 200 ms (Isik et al., 2018). Visual features include the spatial scale of an action (i.e. fine-scale manipulations like knitting versus full-body movements like running) represented throughout visual cortex (Tarhan and Konkle, 2020), and information about biological motion, thought to be extracted within 200 ms in superior temporal cortex (Giese and Poggio, 2003; Hirai et al., 2003; Hirai and Hiraki, 2006; Johansson, 1973; Jokisch et al., 2005; Vangeneugden et al., 2014). Responses in occipito-temporal areas have been shown to reflect semantic features like invariant action category (Hafri et al., 2017; Lingnau and Downing, 2015; Tucciarelli et al., 2019, 2015; Wurm and Caramazza, 2019; Wurm and Lingnau, 2015), as well as social features like the number of agents and sociality of actions (Tarhan and Konkle, 2020; Wurm et al., 2017; Wurm and Caramazza, 2019).

Among the visual, semantic, and social features thought to be processed during action observation, it is unclear which underlie our everyday perception in naturalistic settings. Mounting evidence suggests that naturalistic datasets are key to improving ecological validity and reliability in visual and social neuroscience (Haxby et al., 2020; Nastase et al., 2020; Redcay and Moraczewski, 2020). Most action recognition studies to date have used controlled images and videos showing actions in simple contexts (Isik et al., 2018; Wurm and Caramazza, 2019). However, presenting actions in natural contexts is critical, as stimulus-context interactions have been shown to modulate neural activity (Willems and Peelen, 2021). Recent attempts to understand naturalistic action perception, however, have yielded mixed results, particularly with regard to the role of social features. For example, one recent study concluded that sociality (i.e., presence of a social interaction) was the primary organizing dimension of action representations in the human brain (Tarhan and Konkle, 2020). Another, however, found that semantic action category explained the most variance in fMRI data, with little contribution from social features (Tucciarelli et al., 2019).

Here, we combined a new large-scale dataset of everyday actions with a priori feature labels to comprehensively sample the hypothesis space defined by previous work. This is essential in light of the conflicting results from previous studies, as it allowed us to disentangle the contributions of distinct but correlated feature spaces. We used three-second videos of everyday actions from the "Moments in Time" dataset (Monfort et al., 2019) and replicated our results across two different stimulus sets. Action videos were sampled from different categories based on the American Time Use Survey (ATUS, 2019) and were highly diverse, depicting a variety of contexts and people. We quantified a wide range of visual, action-related, and social-affective features in the videos and, through careful curation, ensured that they were minimally confounded across our dataset.

We used this dataset to probe the behavioral and neural representational space of human action perception. To understand the features that support natural action viewing, we predicted behavioral similarity judgments using the visual, action-related, and social-affective feature sets. Next, to investigate the neural dynamics of action perception, we recorded electroencephalography (EEG) data while participants viewed the stimuli, and we used the three sets of features to predict time-resolved neural patterns.

We found that social-affective features predict action similarity judgments better than, and independently of, visual and action-related features. Visual and action-related features explained less variance in behavior, even though they included fundamental features such as the scene setting and the semantic category of each action. Neural patterns revealed that behaviorally-relevant features are automatically extracted by the brain in a progression from visual to action to social features. Together, our results reveal the importance of social-affective features in how we represent other people's actions, and show that these representations emerge in the brain along a temporal gradient.

Results

Disentangling visual, action, and social-affective features in natural videos

We curated two sets of naturalistic three-second videos of everyday actions from the Moments in Time dataset (Monfort et al., 2019). The videos were selected from a larger set, ensuring that features of interest were minimally correlated (see Supplementary Methods). 18 common activities based on the National Bureau of Labor Statistics' American Time Use Survey (ATUS, 2019) were represented (Table 1; see Methods, section *Behavior: Stimuli*). The two stimulus sets contained 152 videos (8 videos per activity and 8 additional videos with no agents) and 65 videos (3-4 videos per activity) respectively. The second set was used to replicate behavioral results in a separate experiment with different stimuli and participants.

Naturalistic videos of actions can vary along numerous axes, including visual features (e.g. the setting in which the action takes place), action-specific features (e.g. semantic action category), and social-affective features (e.g. the number of agents involved or perceived arousal). To evaluate these different axes, we quantified 17 visual, action-related and social-affective features using image properties, labels assigned by experimenters, and behavioral ratings collected in online experiments (Figure 1a). Visual features ranged from low-level (e.g. pixel values) to high-level (e.g. activations from the final layer of a pretrained neural network). Action-related features included transitivity (object-relatedness), activity (the amount of activity in a video), effectors (body parts involved), and action category based on the American Time Use Survey (ATUS, 2019). Finally, social-affective features included sociality, valence, arousal and number of agents

(see Methods, section *Representational similarity analysis*). Representational dissimilarity matrices (RDM) were created for each feature by calculating pairwise Euclidean distances between all videos.

In both video sets there were only weak correlations between visual features and the higher-level action/social-affective features (Figure 1a). The highest correlations were those within each of the three sets of features, including visual features (Exp 1: Conv1 and image saturation/gist, τ_A=0.29; Exp 2: Conv1 and image hue, τ_A=0.32), action features (Exp 1: arousal and activity, τ_A=0.31; Exp 2: activity and effectors, τ_A=0.33) and social features (sociality and number of agents; Exp 1: τ_A=0.31, Exp 2: τ_A=0.3).

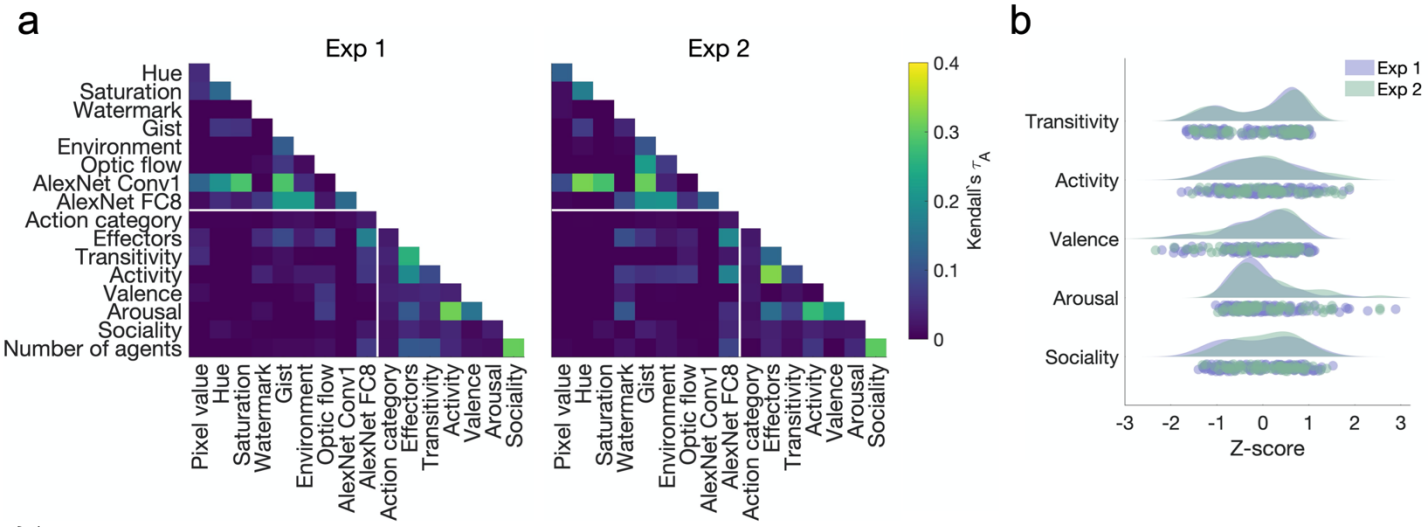


Figure 1. Quantifying visual, social-affective and action features in the two stimulus sets. **a.** Correlations between feature RDMs. Note the low correlations between visual features and action/social-affective features (white rectangle). **b.** Behavioral rating distributions in the two stimulus sets (plots: Allen et al., 2019).

The distributions of action and social-affective features were not significantly different between the two stimulus sets (all Mann-Whitney z<1.08, P>0.28). The width of these distributions suggests that the stimuli spanned a wide range along each feature (Figure 1b). In both experiments, transitivity was notable through its bimodal distribution, likely reflecting the presence or absence of objects in scenes, while other features had largely unimodal distributions.

Individual feature contributions to behavioral similarity

To characterize human action representations, we collected behavioral similarity ratings for all pairs of videos in each set in two multiple arrangement experiments (see Methods, section *Multiple arrangement*). Participants arranged videos according to their similarity inside a circular arena (Figure 2). The task involved arranging different subsets of 3-8 videos until sufficiently reliable distance estimates were reached for all pairs of videos. Videos would play on hover, and participants had to play and move each video to proceed to the next trial. In Experiment 1, participants arranged different subsets of 30 videos out of the total 152, while in Experiment 2, participants arranged all 65 videos. To emphasize natural behavior, participants were not given specific criteria to use when judging similarity. Behavioral RDMs containing the Euclidean distances between all pairs of stimuli were reconstructed from each participant's multiple arrangement data using inverse MDS (Kriegeskorte and Mur, 2012).

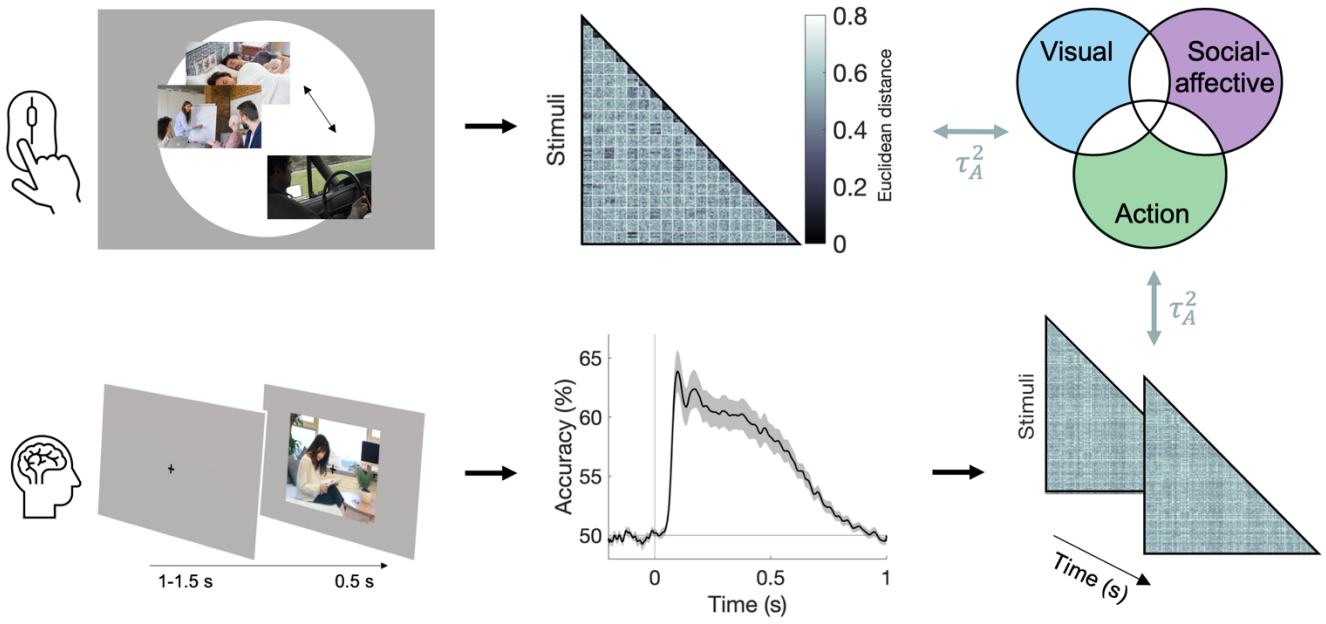


Figure 2. Experimental and analysis pipeline for evaluating the contribution of different features to action representations. Above: a multiple arrangement task was used to generate behavioral RDMs in the two behavioral experiments. Below: EEG data was recorded during a one-back task, and time-resolved neural RDMs were generated using pairwise decoding accuracies. Cross-validated variance partitioning was used to assess the unique contributions of visual, social-affective, and action features to the behavioral and neural RDMs, quantified as the predicted squared Kendall's τ_A. The stimuli in this figure are public domain images similar to the types of videos used in the experiments.

Data reliability was quantified using leave-one-subject-out correlations of the similarity ratings and was above chance in both experiments (Kendall's τ_A = 0.13±0.08 and 0.18±0.08 respectively, both P<0.001, permutation testing; Supplementary Figure 1a). Reliability was significantly higher in Experiment 2 than in Experiment 1 (Mann-Whitney z=3.21, P=0.0013), potentially reflecting differences in both participant pools and sampling methods (subsets of videos in Experiment 1 versus full video dataset in Experiment 2; see Methods, section *Multiple arrangement*).

We assessed the contribution of 17 different visual, social, and action features to behavior in both experiments by correlating each feature RDM to each participant's behavioral RDM (Supplementary Table 2). In Experiment 1 (Figure 3), only two visual features were significantly correlated with the behavioral RDMs (environment and activations from the final fully-connected layer FC8 of AlexNet). However, there were significant correlations between behavioral RDMs and all action-related RDMs (action category, effectors, transitivity and activity), as well as all social-affective RDMs (valence, arousal, sociality and number of agents).

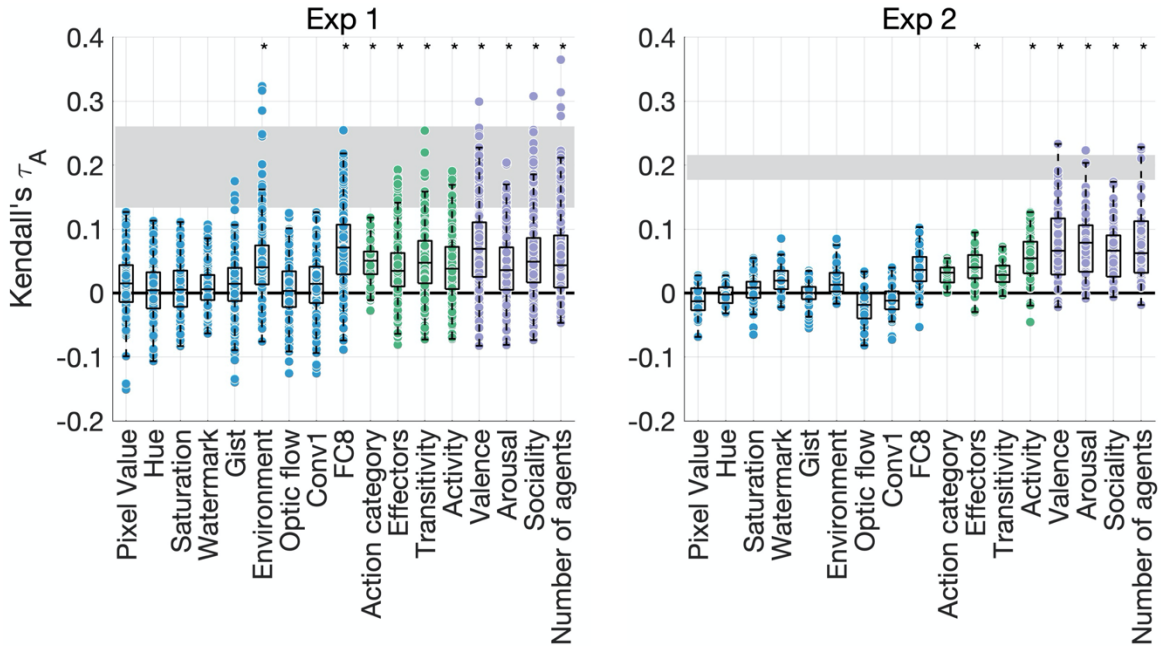
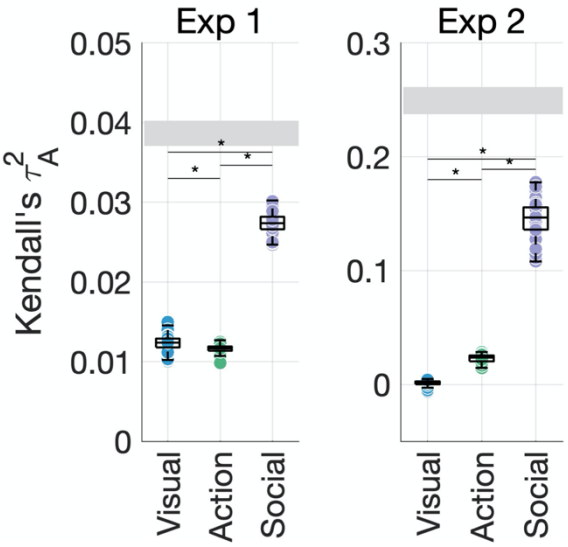


Figure 3. Feature contributions to behavioral similarity. The correlations between each feature RDM and the behavioral RDMs are plotted against the noise ceiling (gray). Each dot is an individual subject. Asterisks denote significance (P<0.005).

145 In Experiment 2, the only visual feature that contributed to behavior was the final fully-
146 connected layer of AlexNet (Figure 3). Among action features, only effectors and activity were
147 significantly correlated with the behavioral RDMs. However, we found significant correlations
148 with all social-affective features. The results thus converge across both experiments in
149 suggesting that social-affective and, to a lesser extent, action-related features, rather than visual
150 properties, explain behavioral similarity.

151 **Social-affective features explain the most unique variance in behavioral representations**

152 We performed a cross-validated variance partitioning analysis (Groen et al., 2018; Lescroart et
153 al., 2015; Tarhan et al., 2021) to determine which features contributed the most *unique* variance
154 to behavior (see Methods, section *Variance partitioning*). We selected the ten features that
155 contributed significantly to behavior in either experiment, i.e. two visual features (environment
156 and layer FC8 of AlexNet) and all action and social-affective features. To keep the analysis
157 tractable and understand the contribution of each type of information, we grouped these features
158 according to their type (visual, action and social-affective) and used them as predictors in a
159 cross-validated hierarchical regression (Figure 4). Note that there was no collinearity among the
160 ten predictors, with an average variance inflation factor of 1.34 (Experiment 1) and 1.37
161 (Experiment 2).



162

Figure 4. Social-affective features explain behavior better than visual and action features. The unique variance explained by visual, action, and social-affective features is plotted against the split-half reliability of the data (gray). Significant differences are marked with asterisks (all $P<0.001$).

Together, the ten predictors explained most of the systematic variance in behavior. In Experiment 1, the predicted squared Kendall's τ_A of the full model was higher on average than the true split-half squared correlation (τ_A^2=0.06±0.001 and τ_A^2=0.04±0.002 respectively). This is likely to be due to the lower reliability of the behavioral similarity data in this experiment, and suggests that the ten predictors are able to explain the data well despite the overall lower prediction accuracy. In Experiment 2, the full model achieved a predicted τ_A^2 of 0.18±0.1 on average, compared to a true squared correlation of 0.25±0.1, suggesting that the ten predictors explain most of the variance (73.21%) in the behavioral data.

In both experiments, social-affective features contributed significantly more unique variance to behavior than visual or action features (Figure 4, all Wilcoxon $z>5.5$, all $P<0.001$). While all three groups of features contributed unique variance to behavior in Experiment 1 (all $P<0.001$, randomization testing), in Experiment 2, only social-affective features contributed significantly to behavior ($P<0.001$), while visual and action features did not ($P=0.06$ and $P=0.47$ respectively). Shared variance between feature groups was not a significant contributor in either dataset.

The semantic RDM included among the action features was a categorical model based on activity categories (ATUS, 2019). To assess whether a more detailed semantic model would explain more variance in behavior, we generated a feature RDM using WordNet similarities between the verb labels corresponding to the videos in the Moments in Time dataset. However, replacing the action category RDM with the WordNet RDM did not increase the variance explained by action features (Supplementary Figure 2).

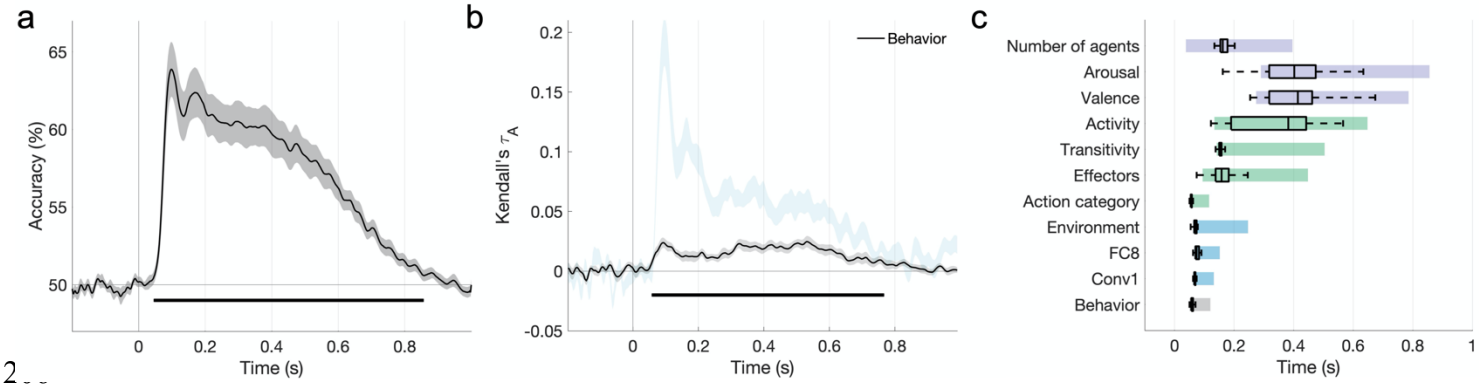
Among the social-affective features we tested, the number of agents could be seen as straddling the visual and social domains. To assess whether our results were driven by this feature, we performed a control variance partitioning analysis pitting the number of agents against the other, higher-level social-affective features (Supplementary Figure 3). In both experiments, the higher-

191 level features (sociality, valence and arousal) contributed more unique variance than the number
192 of agents, suggesting that our results are not explained by purely visual factors.

193

194 **EEG patterns reflect behavioral similarity**

195 We performed an EEG experiment to investigate how action-relevant features are processed over
196 time. Participants viewed 500 ms segments of the 152 videos from Experiment 1 and performed
197 a one-back action task in which they detected repetitions of the action category (see Methods,
198 section *EEG: Experimental procedure*). To relate neural patterns to behavioral and feature
199 RDMs, we computed time-resolved neural RDMs for each participant using decoding accuracies
200 between all pairs of videos (Figures 2, 5a). The time-course of decoding performance was similar
201 to that observed in previous E/MEG studies using still visual stimuli (Carlson et al., 2013; Cichy
202 et al., 2014; Dima et al., 2018; Greene and Hansen, 2018; Isik et al., 2014). Decoding accuracy
203 rose above chance at 50 ms after video onset, reached its maximum at 98 ms (63.88±6.82%
204 accuracy), and remained above chance until 852 ms after video onset (cluster-corrected P<0.05,
205 sign permutation testing).



207 **Figure 5.** The features that explain behavioral action representations also contribute to neural representations. **a.**
208 Time-course of video decoding accuracy, averaged across all pairs of videos and participants (in gray: SEM across
209 participants). The horizontal line marks above-chance performance (sign permutation testing, cluster-corrected
210 P<0.05). **b.** Behavioral similarity correlates with the neural RDM. The noise ceiling is shown in light blue (leave-
211 one-subject out correlation, mean±SD). Horizontal lines mark significant time windows (P<0.05, cluster-corrected).
212 **c.** The distribution of significant correlation onsets for each feature model across 1000 bootstrapping iterations
213 (P<0.05, cluster-corrected). Color rectangles show 90% confidence intervals.

To assess brain-behavior correlations, we related the average behavioral RDM obtained in Experiment 1 to the time-resolved neural RDMs (Kendall's τ_A). The behavioral RDM correlated significantly with neural patterns during a cluster between 62 and 766 ms after video onset (Figure 5b), suggesting that the features guiding the intuitive categorization of naturalistic actions also underlie their neural organization.

Neural timescale of individual feature representations

We assessed the correlations between EEG patterns and the ten feature RDMs found to contribute to behavior in Experiment 1. We also included an additional feature RDM based on the first convolutional layer of AlexNet, which best captures early visual neural responses (Supplementary Figure 7; see Methods, section *Multivariate analysis*). The feature RDMs that contributed to behavioral similarity also correlated with the EEG patterns (Supplementary Figures 7-9), with a single exception (sociality).

A bootstrapping analysis of the cluster onsets of these correlations (Figure 5c) suggests a progression from visual to action and social-affective features. Visual predictors correlated with the neural patterns between 65±15 ms (mean ± SD, Conv1) and 84±62 ms (Environment), while action category also had an early onset (58±9 ms). Other action-related features, however, emerged later (transitivity: 170±67 ms, effectors: 192±94 ms, activity: 345±133 ms). Among social-affective features, the number of agents had the earliest correlation onset (178±81 ms), while valence and arousal emerged later (395±81 and 404±91 ms respectively). Importantly, these features are spontaneously extracted in the brain, as none of them, with the exception of action category, were directly probed in the one-back task performed by participants. In addition, all features were extracted during behaviorally relevant time windows (Figure 5b).

A temporal hierarchy in action perception

A cross-validated variance partitioning analysis revealed different stages in the processing of naturalistic actions (Figure 6). Visual features dominated the early time windows (66-138 ms after video onset). Action features also contributed a significant amount of unique variance (162-598 ms), as well as variance shared with social-affective features (354-598 ms; Supplementary Figure 5). Finally, social-affective features independently predicted late neural responses (446-

782 ms). Importantly, visual features did not share a significant amount of variance with either action or social-affective features.

An analysis of effect onsets across 100 split-half iterations points to the hierarchical processing of these features, with a progression from visual to action to social-affective features. Social-affective features (mean onset 418±89 ms) contributed unique variance significantly later than other feature sets, while action features (245±104 ms) came online later than visual features (65±8 ms; all Wilcoxon z>7.27, P<0.001; Figure 6b). A fixed-effects analysis revealed the same order of feature information with larger effect sizes (Supplementary Figure 6).

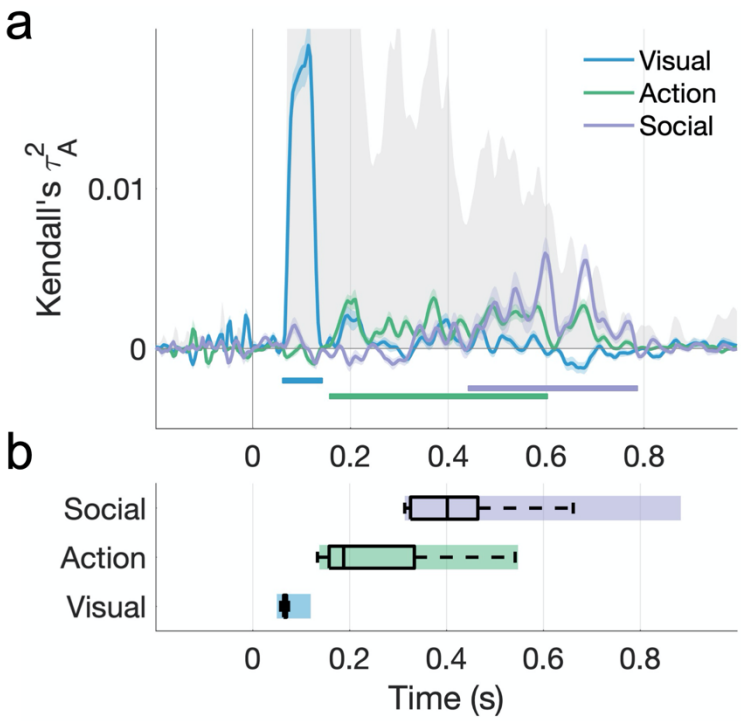


Figure 6. Hierarchical processing of visual, action, and social-affective features. **a.** Unique variance explained by each group of features over time. The split-half reliability of the data is shown in gray (shaded area; see also Figure 5b). **b.** The distribution of effect onsets across 100 split-half iterations. Color rectangles show 90% confidence intervals.

Discussion

Here, we used a large-scale naturalistic stimulus set to disentangle the roles of different features in action perception. Two novel findings emerge from our study. First, our behavioral results

suggest that social-affective features play the most important role in how we organize naturalistic everyday actions, above and beyond fundamental visual and action features like scene setting or action category. Second, these behaviorally-relevant features are spontaneously extracted in the brain, and follow a hierarchical sequence from visual to action-related and culminating with social-affective features. These results offer an account of how internal representations of everyday actions emerge in the mind and brain.

Behavioral representations: what features support action perception?

Across two separate multiple arrangement experiments with large-scale naturalistic stimulus sets, we found that social-affective features predicted similarity judgments better than, and independently of, visual and action-related features. By sampling a comprehensive feature space ranging from low-level to conceptual, we were able to distinguish between components that often covary, such as scene setting and action category or sociality and transitivity. Previous studies have operationalized features in different ways, and an exhaustive investigation is thus difficult; however, our approach of including several important features from each group mitigated this, as suggested by the high amount of variance in behavior collectively explained by our features.

Our work adds to a growing body of evidence for the importance of social features in action perception, and extends it by disentangling the contributions of specific social and semantic features. Previous work has highlighted sociality as an essential feature in neural action representations (Tarhan and Konkle, 2020; Wurm et al., 2017; Wurm and Caramazza, 2019) and a recent study (Tarhan et al., 2021) found that behavioral action similarity judgments were better explained by similarity in actors' goals than by visual similarity. In line with this work, we found a minimal contribution of visual features to action similarity judgments. In contrast, all of our social-affective features – the number of agents, sociality, valence, and arousal – were significantly correlated with behavioral similarity. Furthermore, only two individual action-related features replicated across the two experiments: the amount of activity and the effector (body part) feature, the latter of which is highly relevant to the actors' goals. This could be interpreted as further evidence for the importance of socially-relevant features in our internal

representations of actions, and identifies specific social and goal-related features that are important for action understanding.

A hypothesis-driven approach will always pose challenges, due to practical limitations in the number of feature spaces one can feasibly test. Our approach of grouping predictors together based on theoretical distinctions made it possible to rigorously evaluate the unique contributions of different types of features, which is an essential first step in understanding naturalistic action representations. This analysis revealed that social-affective features contributed the most unique variance in both experiments, suggesting that they robustly predict behavioral similarity judgments, while visual and action features explained little unique variance in either experiment (Figure 4).

Our social-affective feature space included one feature that could be construed as a perceptual precursor to sociality, namely the number of agents in each video. Indeed, previous fMRI work has suggested that neural representations of actions in the visual system reflect perceptual precursors of social features, rather than higher-level social features (Wurm and Caramazza, 2019). Here, we found that high-level social-affective features (sociality, valence and arousal) contributed significantly to behavior independently of the number of agents. Further, these high-level social-affective features explained significantly more unique variance in behavior than the number of agents in both experiments (Supplementary Figure 3). Our findings suggest that high-level social-affective features uniquely drive human action representations.

Neural representations: how does action perception unfold over time?

Using EEG, we tracked the temporal dynamics of naturalistic action perception. Using naturalistic stimuli and a rich feature space enabled us to disentangle the contributions of different features and investigate their relative timing. Visual, action, and social-affective features made unique contributions to the EEG patterns at different processing stages, revealing a representational hierarchy of spontaneously-extracted features.

Almost all behaviorally-relevant features correlated with the EEG patterns, with action-related and social-affective features emerging later than visual features (Figure 5c). Most action-related features emerged within 200 ms, on the timescale of feedforward processing, which is consistent with prior work

showing invariant responses to actions as early as 200 ms (Isik et al., 2018; Tucciarelli et al., 2015), and action transitivity processing as early as 250 ms (Wamain et al., 2014). Among social-affective features, the number of agents emerged earliest (162 ms), pointing to the role of this feature as a perceptual precursor in social perception (Papeo, 2020; Wurm and Caramazza, 2019). Valence and arousal emerged later, around 400 ms after video onset. Interestingly, sociality, which has been highlighted as an important dimension in previous fMRI work on action perception (Tarhan and Konkle, 2020; Wurm et al., 2017), did not correlate with the EEG patterns. While the absence of an effect does not preclude the possibility that this feature is being processed, it is possible that prior work has confounded sociality with other correlated social-affective features (such as the number of agents, or arousal).

Variance partitioning revealed a clear temporal progression from visual features (~100 ms) to action features (~150-600 ms) to social-affective features (~400-600 ms). Importantly, these processing stages emerged after partialling out the contributions of other groups of predictors in a cross-validated analysis, validating our a priori distinctions between feature classes. These findings suggest that the extraction of visual features occurs rapidly, within 200 ms, and is likely supported by feedforward computations. The social-affective features that support behavioral representations, however, were extracted last. This is consistent with theories suggesting that internal visual experience reverses the stages of perceptual processing (Dijkstra et al., 2020; Hochstein and Ahissar, 2002). Specifically, it was the final, social-affective stage of neural processing that was reflected in the intuitive behavioral representations, and not the initially extracted visual features. Furthermore, action-related features were extracted significantly before social-affective features, suggesting the two are not extracted in parallel, but instead pointing to a hierarchy in which both visual and action-related features may contribute to socially relevant computations.

Our results add temporal characterization to previous fMRI findings, suggesting that the seemingly conflicting features revealed by previous studies, like sociality (Tarhan and Konkle, 2020) or semantic action category (Tucciarelli et al., 2019), emerge at different stages during action observation. Thus, the existence of different organizing dimensions can be explained not just through spatial segregation within and across brain areas, but also through a temporal gradient starting with visual features and concluding with behaviorally-relevant social representations. More work is needed to understand where these dynamic representations emerge

in the brain, and whether they are supported by overlapping or distinct networks. Future research could test this using EEG-fMRI fusion to track the spatiotemporal dynamics of action representations.

Actions in context

As real-world actions tend to occur in a rich social context, studies of action perception should consider social features and the interactions between different systems for perceiving actions, agents and their mental states (Quadflieg and Koldewyn, 2017). Recent works suggests that social perception enhances visual processing (Bellot et al., 2021; Papeo, 2020) and recruits dedicated neural circuits (Isik et al., 2017; Pitcher and Ungerleider, 2021). Our findings open exciting new avenues for connecting these areas of research. For example, future studies could more explicitly disentangle the perceptual and conceptual building blocks of social and affective features, such as body posture or facial expression, and their roles in action and interaction perception.

One fundamental question that lies at the root of this work is how actions should be defined and studied. Here, we adopted a broad definition of the term, focusing on activities as described in the American Use Survey (ATUS, 2019). Although our stimuli were selected to clearly depict short, continuous actions performed by visible agents, their naturalistic and context-rich nature means that they could be understood as “events”, encompassing elements that are not singularly specific to actions. A wealth of evidence has shown that context changes visual processing in a non-additive way (Bar, 2004; Willems and Peelen, 2021), and emerging evidence suggests that the same is true for actions (Wurm et al., 2012). Studying actions in context holds promise for understanding how semantically rich representations emerge in naturalistic vision. This, in turn, will pave the way towards a computational understanding of the neural processes that link perception and cognition.

Methods

Behavior: Stimuli

We curated two stimulus sets containing three-second videos of everyday actions from the Moments in Time dataset (Monfort et al., 2019). To broadly sample the space of everyday actions, we first identified the most common activities from the National Bureau of Labor Statistics' American Time Use Survey (ATUS, 2019). We adjusted this list to include both social and non-social activities that lend themselves to visual representation (Table 1). In particular, "educational activities" were included both under "working" and a more specific "instructing" category. In addition, some broad categories ("leisure and sports", "household activities") were split into more specific ones. Finally, we added a "fighting" category to capture unpleasant social interactions, and a "driving" category (as travel was only included under "other activities") .

Activity	**Verb labels**
Childcare / taking care of children	crying, cuddling, feeding, giggling, socializing
Driving	driving, socializing
Eating	chewing, eating
Fighting	fighting
Gardening	gardening, mowing, planting, shoveling, weeding
Grooming	bathing, brushing, combing, trimming, washing
Hiking	hiking
Housework	cleaning, dusting, repairing, scrubbing, vacuuming
Instructing	instructing, teaching
Playing games	gambling, playing+fun, playing+videogames, socializing
Preparing food	barbecuing, boiling, chopping, cooking, frying, grilling, rinsing, stirring
Reading	reading
Religious activities	praying, preaching
Sleeping	resting, sleeping
Socializing and social events	celebrating, dancing, marrying, singing, socializing, talking
Sports	exercising, playing+sports, swimming, throwing
Telephoning	calling, telephoning
Working	working
Control videos	blowing, floating, raining, shaking

Table 1. Activities from the ATUS included in each of the two stimulus sets, with the corresponding verb labels from the Moments in Time dataset. Note that control videos were only included in the first dataset.

We curated an initial set of approximately 500 videos from the Moments in Time dataset (see Supplementary Methods) by identifying the verb labels relevant to our chosen activities. We then selected two subsets of videos (1) that sampled all activities in a balanced manner, and (2) where sociality (as assessed through behavioral ratings, see below Section 2.3) was minimally correlated to the number of agents (experimenter-labeled). For more details see Supplementary Methods. These two features are difficult to disentangle in naturalistic stimulus sets, and we were able to minimize, though not fully eliminate, this correlation (Figure 1a).

The first stimulus set contained 152 videos (8 videos per activity and 8 additional videos with no agents) and was used in Experiment 1. The second stimulus set contained 65 videos (3-4 videos per activity) and was used in Experiment 2. The videos were preprocessed to a framerate of 30 frames-per-second and resized to 600 x 400 pixels.

Behavior: Participants

Behavioral ratings

A total of 256 workers (202 after exclusions, located in the United States, worker age and gender not recorded) from the online platform Amazon Mechanical Turk provided sociality, valence, arousal, and activity ratings of the video stimuli, and 43 workers (35 after exclusions) provided transitivity ratings.

Multiple arrangement

Two separate online multiple arrangement experiments were performed on each of the two stimulus sets. A total of 374 workers from Amazon Mechanical Turk took part in Experiment 1 (300 after exclusions, located in the United States, worker age and gender not recorded). Experiment 2 involved 58 participants (53 after exclusions, 31 female, 20 male, 1 non-binary, 1 not reported, mean age 19.38±1.09) recruited through the Department of Psychological and Brain Sciences Research Portal at Johns Hopkins University.

All procedures for online data collection were approved by the Johns Hopkins University Institutional Review Board and informed consent was obtained from all participants.

Behavior: Experimental procedure

Behavioral ratings

Participants viewed subsets of 30-60 videos from the initially curated large-scale set and rated the events depicted on a five-point scale. In a first set of experiments, the dimensions rated were: sociality (how social the events were, from 1 – not at all to 5 – very social); valence (how pleasant the events were, from 1 – very unpleasant to 5 – very pleasant); arousal (how intense the events were, from 1 – very calm to 5 – very intense); and activity (how active they were, from 1 – no action to 5 – very active). In separate experiments, participants provided transitivity ratings for the two final stimulus sets (i.e. to what extent the actions involved a person or people interacting with an object, from 1 – not at all to 5 – very much). This amounted to an average of 17.46±2.14 ratings per video (Experiment 1) and 18.22±2.09 ratings per video (Experiment 2). The experiments were implemented in JavaScript using the jsPsych library (de Leeuw, 2015).

Multiple arrangement

To characterize human action representations, we collected behavioral similarity ratings using two multiple arrangement experiments. The experiments were conducted on the Meadows platform (www.meadows-research.com) and required participants to arrange the videos according to their similarity inside a circular arena. Participants were free to use their own criteria to determine similarity, so as to encourage natural behavior.

Each trial started with the videos arranged around the circular arena. The videos would start playing on hover, and the trial would not end until all videos were played and dragged-and-dropped inside the arena (Figure 2). Different sets of videos were presented in different trials. An adaptive "lift-the-weakest" algorithm was used to resample the video pairs placed closest together, so as to gather sufficient evidence (or improve the signal-to-noise ratio) for each pair. This procedure was repeated until an evidence criterion of 0.5 was reached for each pair, or until the experiment timed out (Experiment 1: 90 minutes, Experiment 2: 120 minutes). By asking participants to zoom into the subsets previously judged as similar, the task required the use of different contexts and criteria to judge relative similarities. Compared to other methods of measuring similarity, multiple arrangement thus combines efficient sampling of a large stimulus

set with adaptive behavior that can recover a multi-dimensional similarity structure (Kriegeskorte and Mur, 2012).

In Experiment 1, participants arranged different subsets of 30 videos from the 152-video set, with a maximum of 7 videos shown in any one trial. The stimuli were sampled in a balanced manner across participants. The task took on average 32±14.4 minutes and 86.8±22.6 trials.

In Experiment 2, all participants arranged the same 65 videos (entire 65-video set), with a maximum of 8 videos shown in any one trial. The task took on average 87.5±24.6 minutes, including breaks, and 289.7±57.3 trials.

Both experiments included a training trial in which participants arranged the same 7 videos before beginning the main task. Participants were excluded from further analysis if there was a low correlation between their training data and the average of all other participants' data (over 2 standard deviations below the mean). They were also excluded if they responded incorrectly to a catch trial requiring them to label the action in previously seen videos.

Inverse MDS was used to construct behavioral dissimilarity matrices containing normalized Euclidean distances between all pairs of videos (Kriegeskorte and Mur, 2012). In Experiment 1, the behavioral RDM contained 11476 pairs with an average of 11.37±3.08 ratings per pair; in Experiment 2, there were 2080 pairs rated by all 53 participants.

Behavior: Data analysis

Representational similarity analysis

Everyday actions can be differentiated along numerous axes. Perceptually, they can differ in terms of visual properties, like the setting in which they take place. They can also be characterized through action-related features like semantic action category, or through social features, like the number of agents involved. Understanding how these features contribute to natural behavior can shed light on how naturalistic action representations are organized. Here, we used representational similarity analysis (RSA) to assess the contribution of visual, action, and social-affective features to the behavioral similarity data.

We quantified features of interest using image properties, labels assigned by experimenters, and behavioral ratings (provided by participants, see above Section 2.3). We calculated the Euclidean distances between all pairs of stimuli in each feature space, thus generating 17 feature RDMs.

To quantify visual features, image properties were extracted separately for each frame of each video and averaged across frames. These included pixel value, hue, saturation, motion energy (the magnitude of the optic flow estimated using the Horn-Schunck method), and the spatial envelope of each image quantified using GIST (Oliva and Torralba, 2001). We also extracted activations from the first convolutional layer and last fully-connected layer of a pre-trained feedforward convolutional neural network (AlexNet; Krizhevsky et al., 2012). Two additional experimenter-labeled features were included: scene setting (indoors/outdoors) and the presence of a watermark.

Action feature RDMs were based on transitivity and activity ratings (provided by participants, see above), as well as action category (a binary RDM clustering the stimuli into activity categories based on the initial dataset designations) and effectors (experimenter-labeled). The latter consisted of binary vectors indicating the involvement of body parts in each action (face/head, hands, arms, legs, and torso). To assess whether a more detailed semantic model would capture more information, we also performed a control analysis using a feature RDM based on WordNet similarities between the verb labels in the "Moments in Time" dataset (Supplementary Figure 2).

Social-affective feature RDMs were based on sociality, valence, and arousal ratings (all provided by participants, see Behavioral Ratings above) and the number of agents in each video, which was labeled by experimenters on a four-point scale (from 0, no agent present, to 3, three or more agents present).

Each subject's behavioral RDM was correlated to the feature RDMs, and the resulting Kendall's τ_A values were tested against chance using one-tailed sign permutation testing (5000 iterations). P-values were omnibus-corrected for multiple comparisons using a maximum correlation threshold across all models (Nichols and Holmes, 2001).

A noise ceiling was calculated by correlating each subject's RDM to the average RDM (upper bound), as well as to the average RDM excluding the left-out subject (lower bound; Nili et al., 2014).

Variance partitioning

Despite low correlations between features of interest in both stimulus sets (Figure 1a), shared variance could still contribute to the representational similarity analysis results. To estimate the unique contributions of the three primary groups of features, we performed a cross-validated variance partitioning analysis, excluding individual features that did not correlate with the behavioral data in the above RSA analysis. The three groups included: visual features (scene setting and the last fully-connected layer of AlexNet), action features (action category, effectors, transitivity, action), and social-affective features (number of agents, sociality, valence, arousal).

The behavioral data were randomly split into training and test sets (100 iterations) by leaving out half of the individual ratings for each pair of videos in Experiment 1 (since different participants saw different subsets of videos) or half of the participants in Experiment 2. We fit seven different regression models using the average training RDM (with every possible combination of the three groups of features), and we calculated the squared Kendall's τ_A between the predicted responses and the average test RDM. These values were then used to calculate the unique and shared portions of variance contributed by the predictors (Groen et al., 2018; Lescroart et al., 2015; Tarhan et al., 2021).

The resulting values were tested against chance using one-tailed sign permutation testing (5000 iterations, omnibus-corrected for multiple comparisons). Differences between groups of features were assessed with two-sided Wilcoxon signed-rank tests.

EEG: Stimuli

The stimulus set from behavioral Experiment 1 was used in the EEG experiment, containing 152 videos from 18 categories, as well as control videos. The three-second stimuli were trimmed to a duration of 0.5 seconds centered around the action to improve time-locking to the EEG signals and allow for a condition-rich experimental design. An additional 50 videos were included as catch stimuli (25 pairs depicting the same action, manually chosen from the larger stimulus set).

EEG: Participants

Fifteen participants (6 female, 9 male, mean age 25.13±6.81) took part in the EEG experiment. All participants were right-handed and had normal or corrected-to-normal vision. Informed consent was obtained in accordance with the Declaration of Helsinki and all procedures were approved by the Johns Hopkins University Institutional Review Board.

EEG: Experimental procedure

Continuous EEG recordings with a sampling rate of 1000 Hz were made with a 64-channel Brain Products ActiCHamp system using actiCAP electrode caps in a Faraday chamber. Electrode impedances were kept below 25 kΩ when possible and the Cz electrode was used as an online reference.

Participants were seated upright while viewing the videos on a back-projector screen situated approximately 45 cm away. The 152 videos were shown in pseudorandom order in each of 10 blocks with no consecutive repetition allowed. In addition, four repetitions of the 25 catch video pairs were presented at random times during the experiment, with the pairs shuffled to minimize learning effects. Participants performed a one-back task and were asked to press a button on a Logitech game controller when they detected two consecutive videos showing the same action. There was a break every 150 trials and participants could continue the experiment by pressing a button. In total, the experiment consisted of 1720 trials (1520 experimental trials and 200 catch trials) and took approximately 45 minutes.

Each trial started with a black fixation cross presented on a gray screen for a duration chosen from a uniform distribution between 1 and 1.5 s, followed by a 0.5 s video. The stimuli were presented on the same gray background and subtended approximately 15 x 13 degrees of visual angle. The fixation cross remained on screen and participants were asked to fixate throughout the experiment. A photodiode was used to accurately track on-screen stimulus presentation times and account for projector lag. The paradigm was implemented in MATLAB R2019a using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

EEG: Data analysis

Preprocessing

EEG data preprocessing was performed using MATLAB R2020b and the FieldTrip toolbox (Oostenveld et al., 2011). First, the EEG data were aligned to stimulus onset using the photodiode data to correct for any lag between stimulus triggers and on-screen presentation. The aligned data were segmented into 1.2 s epochs (0.2 s pre-stimulus to 1 s post-stimulus onset), baseline-corrected using the 0.2 s prior to stimulus onset, and high-pass filtered at 0.1 Hz.

Artefact rejection was performed using a semi-automated pipeline. First, the data were filtered between 110 and 140 Hz and Hilbert-transformed to detect muscle artefacts; segments with a z-value cutoff above 15 were removed. Next, channels and trials with high variance were manually rejected based on visual inspection of a summary plot generated using the *ft_rejectvisual* function in FieldTrip. Finally, independent component analysis (ICA) was performed to identify and remove eye movement components from the data.

Catch trials were removed from the data together with any trials that elicited a button response (13.74%±1.82% of all trials). Of the remaining trials, 8.36%±5.01% (ranging between 25 and 275 trials) were removed during the artefact rejection procedure. A maximum of two noisy electrodes were removed from eight participants' datasets.

Prior to further analysis, the data were re-referenced to the median across all electrodes, low-pass filtered at 30 Hz to investigate evoked responses and downsampled to 500 Hz.

Multivariate analysis

We performed multivariate analyses to investigate (1) whether EEG patterns reflected behavioral similarity, and (2) whether different visual, action, and social-affective features explained variance in the neural data.

First, time-resolved decoding of every pair of videos was performed using a linear Support Vector Machine classifier as implemented in the LibSVM library (Chang and Lin, 2011). For each pair of videos, pseudotrials were created by splitting each participant's single-trial data into

two subsets and averaging the trials in each subset to improve SNR (Isik et al., 2018). The two pseudotrials were used to train and test the classifier separately at each timepoint, and multivariate noise normalization was performed using the covariance matrix of the training data (Guggenmos et al., 2018). This procedure was repeated 10 times with different data splits. The average decoding accuracies between all pairs of videos were used to generate a time-resolved neural RDM for each participant, with decoding accuracy being used as a measure of dissimilarity.

Next, we evaluated the correlations between each participant's neural RDM and the feature RDMs found to correlate with behavioral similarity (Experiment 1). To investigate the link between behavioral and neural representations, we also correlated neural RDMs with the average behavioral RDM obtained from the multiple arrangement task in Experiment 1. This analysis was performed using 10 ms sliding windows with an overlap of 6 ms. The resulting Kendall's τ_A values were tested against chance using one-tailed sign permutation testing (5000 iterations, cluster-corrected for multiple comparisons across time using the maximum cluster sum, α = 0.05, cluster-setting α = 0.05). A noise ceiling was calculated using the same procedure as in the behavioral RSA (see above Section 2.4). Effect latencies were assessed by bootstrapping the individual correlations 1000 times with replacement to calculate 90% confidence intervals around effect onsets.

To quantify the contributions of visual, social-affective, and action features to the neural RDMs, a time-resolved cross-validated variance partitioning procedure was performed. Using 100 split-half cross-validation iterations, the neural RDM was entered as a response variable in a hierarchical regression with three groups of feature RDMs (visual, social-affective, and action) as predictors. This analysis employed the same ten feature RDMs used in the behavioral variance partitioning (see above Variance partitioning), with the addition of activations from the first convolutional layer of AlexNet (Conv1). As Conv1 best captures early visual responses (Supplementary Figure 7), its inclusion ensured that we did not underestimate the role of visual features in explaining neural variance.

The analysis was carried out using 10 ms sliding windows with an overlap of 6 ms. The resulting predicted Kendall's τ_A values were tested against chance using one-tailed sign permutation

598 testing (5000 iterations, cluster-corrected for multiple comparisons using the maximum cluster
599 sum across time windows and regressions performed, $\alpha = 0.05$, cluster-setting $\alpha = 0.05$). The
600 distributions of effect onsets across the 100 split-half iterations were compared using two-sided
601 Wilcoxon signed-rank tests.

602 Data availability

603 Behavioral and EEG data and results have been archived as an Open Science Framework
604 repository (https://osf.io/hrmxn/). Analysis code is available on GitHub
605 (https://github.com/dianadima/mot_action).

606 Acknowledgments

607 This material is based upon work supported by the Center for Brains, Minds and Machines
608 (CBMM), funded by NSF STC award CCF-1231216. The authors wish to thank Tara Ghazi,
609 Seah Chang, Alyssandra Valenzuela, Melody Lee, Cora Mentor Roy, Haemy Lee Masson, and
610 Lucy Chang for their help with the EEG data collection, Dimitrios Pantazis for pairwise
611 decoding code, and Emalie McMahon for comments on the manuscript.
612

613 References

614 Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: A
615 multi-platform tool for robust data visualization. Wellcome Open Res. 4.
616 https://doi.org/10.12688/wellcomeopenres.15191.1

617 ATUS, 2019. American Time Use Survey, United States Department of Labor. Bureau of Labor
618 Statistics. Washington, DC.

619 Bar, M., 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 617–629.
620 https://doi.org/10.1038/nrn1476

621 Bedny, M., Caramazza, A., 2011. Perception, action, and word meanings in the human brain:
622 The case from action verbs. Ann. N. Y. Acad. Sci. https://doi.org/10.1111/j.1749-

6632.2011.06013.x

Bellot, E., Abassi, E., Papeo, L., 2021. Moving Toward versus Away from Another: How Body Motion Direction Changes the Representation of Bodies and Actions in the Visual Cortex. Cereb. Cortex 1–16. https://doi.org/10.1093/cercor/bhaa382

Brainard, D.H., 1997. The Psychophysics Toolbox. Spat. Vis. 10, 433–436. https://doi.org/http://dx.doi.org/10.1163/156856897X00357

Carlson, T., Tovar, D.A., Alink, A., Kriegeskorte, N., 2013. Representational dynamics of object vision: The first 1000 ms. J. Vis. 13, 1–19. https://doi.org/10.1167/13.10.1

Chang, C.C., Lin, C.J., 2011. LIBSVM: A Library for support vector machines. ACM Trans. Intell. Syst. Technol. 2. https://doi.org/10.1145/1961189.1961199

Cichy, R.M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time. Nat. Publ. Gr. 17, 455–462. https://doi.org/10.1038/nn.3635

de Leeuw, J.R., 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behav. Res. Methods 47, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dijkstra, N., Ambrogioni, L., Vidaurre, D., van Gerven, M., 2020. Neural dynamics of perceptual inference and its reversal during imagery. Elife 9, 1–19. https://doi.org/10.7554/eLife.53588

Dima, D.C., Perry, G., Messaritaki, E., Zhang, J., Singh, K.D., 2018. Spatiotemporal dynamics in human visual cortex rapidly encode the emotional content of faces. Hum. Brain Mapp. 39, 3993–4006. https://doi.org/10.1002/hbm.24226

Giese, M.A., Poggio, T., 2003. Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. 4, 179–192. https://doi.org/10.1038/nrn1057

Greene, M.R., Hansen, B.C., 2018. Shared spatiotemporal category representations in biological and artificial deep neural networks. PLOS Comput. Biol. 14, e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.I., 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7, e32962. https://doi.org/10.7554/eLife.32962

Guggenmos, M., Sterzer, P., Cichy, R.M., 2018. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. Neuroimage 173, 434–447. https://doi.org/10.1016/J.NEUROIMAGE.2018.02.044

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. https://doi.org/10.1523/JNEUROSCI.2496-16.2017

Haxby, J. V., Gobbini, M.I., Nastase, S.A., 2020. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. Neuroimage 216, 116561. https://doi.org/10.1016/j.neuroimage.2020.116561

Hirai, M., Fukushima, H., Hiraki, K., 2003. An event-related potentials study of biological motion perception in humans. Neurosci. Lett. 344, 41–44. https://doi.org/10.1016/S0304-3940(03)00413-0

Hirai, M., Hiraki, K., 2006. The relative importance of spatial versus temporal structure in the perception of biological motion: An event-related potential study. Cognition 99, 15–29. https://doi.org/10.1016/j.cognition.2005.05.003

Hochstein, S., Ahissar, M., 2002. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. Neuron 36, 791–804.

Humphreys, G.F., Newling, K., Jennings, C., Gennari, S.P., 2013. Motion and actions in language: Semantic representations in occipito-temporal cortex. Brain Lang. 125, 94–105. https://doi.org/10.1016/j.bandl.2013.01.008

Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N., 2017. Perceiving social interactions in the posterior superior temporal sulcus. Proc. Natl. Acad. Sci. U. S. A. 114, E9145–E9152. https://doi.org/10.1073/pnas.1714471114

674 Isik, L., Meyers, E.M., Leibo, J.Z., Poggio, T., 2014. The dynamics of invariant object
675 recognition in the human visual system. J. Neurophysiol. 111, 91–102.
676 https://doi.org/10.1152/jn.00394.2013

677 Isik, L., Tacchetti, A., Poggio, T., 2018. A fast, invariant representation for human action in the
678 visual system. J. Neurophysiol. 119, 631–640. https://doi.org/10.1152/jn.00642.2017

679 Jamali, M., Grannan, B.L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., Williams, Z.M., 2021.
680 Single-neuronal predictions of others' beliefs in humans. Nature 591, 610–614.
681 https://doi.org/10.1038/s41586-021-03184-0

682 Johansson, G., 1973. Visual perception of biological motion and a model for its analysis.
683 Percept. Psychophys. 14, 201–211. https://doi.org/10.3758/BF03212378

684 Jokisch, D., Daum, I., Suchan, B., Troje, N.F., 2005. Structural encoding and recognition of
685 biological motion: Evidence from event-related potentials and source analysis. Behav. Brain
686 Res. 157, 195–204. https://doi.org/10.1016/j.bbr.2004.06.025

687 Kleiner, M., Brainard, D.H., Pelli, D.G., Broussard, C., Wolf, T., Niehorster, D., 2007. What's
688 new in Psychtoolbox-3? Perception 36, S14. https://doi.org/10.1068/v070821

689 Kriegeskorte, N., Mur, M., 2012. Inverse MDS: Inferring dissimilarity structure from multiple
690 item arrangements. Front. Psychol. 3, 1–13. https://doi.org/10.3389/fpsyg.2012.00245

691 Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep
692 Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 1–9.
693 https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

694 Lescroart, M.D., Stansbury, D.E., Gallant, J.L., 2015. Fourier power, subjective distance, and
695 object categories all provide plausible models of BOLD responses in scene-selective visual
696 areas. Front. Comput. Neurosci. 9, 1–20. https://doi.org/10.3389/fncom.2015.00135

697 Lingnau, A., Downing, P.E., 2015. The lateral occipitotemporal cortex in action. Trends Cogn.
698 Sci. 19, 268–277. https://doi.org/10.1016/j.tics.2015.03.006

699 Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan,
700 Q., Gutfruend, D., Vondrick, C., Oliva, A., 2019. Moments in Time Dataset: one million
701 videos for event understanding. IEEE Trans. Pattern Anal. Mach. Intell. 1–8.

702 Nastase, S.A., Goldstein, A., Hasson, U., 2020. Keep it real: rethinking the primacy of
703 experimental control in cognitive neuroscience. Neuroimage 222, 117254.
704 https://doi.org/10.1016/j.neuroimage.2020.117254

705 Nichols, T.E., Holmes, A.P., 2001. Nonparametric Permutation Tests For Functional
706 Neuroimaging : A Primer with Examples. Hum. Brain Mapp. 25, 1–25.
707 https://doi.org/10.1002/hbm.1058

708 Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A
709 Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. 10, e1003553.
710 https://doi.org/10.1371/journal.pcbi.1003553

711 Oliva, A., Torralba, A., 2001. Modeling the Shape of the Scene : A Holistic Representation of
712 the Spatial Envelope. Int. J. Comput. Vis. 42, 145–175.
713 https://doi.org/10.1023/A:1011139631724

714 Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: Open source software for
715 advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell.
716 Neurosci. 2011, 156869. https://doi.org/10.1155/2011/156869

717 Papeo, L., 2020. Twos in human visual perception. Cortex 132, 473–478.
718 https://doi.org/10.1016/j.cortex.2020.06.005

719 Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: Transforming numbers
720 into movies. Spat. Vis. 10, 437–442. https://doi.org/10.1163/156856897X00366

721 Pitcher, D., Ungerleider, L.G., 2021. Evidence for a Third Visual Pathway Specialized for Social
722 Perception. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2020.11.006

723 Quadflieg, S., Koldewyn, K., 2017. The neuroscience of people watching: How the human brain
724 makes sense of other people's encounters. Ann. N. Y. Acad. Sci. 1396, 166–182.

https://doi.org/10.1111/nyas.13331

Redcay, E., Moraczewski, D., 2020. Social cognition in context: A naturalistic imaging approach. Neuroimage 216, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Spunt, R.P., Satpute, A.B., Lieberman, M.D., 2011. Identifying the what, why, and how of an observed action: An fMRI study of mentalizing and mechanizing during action observation. J. Cogn. Neurosci. 23, 63–74. https://doi.org/10.1162/jocn.2010.21446

Tarhan, L., De Freitas, J., Konkle, T., 2021. Behavioral and Neural Representations en route to Intuitive Action Understanding 1–22.

Tarhan, L., Konkle, T., 2020. Sociality and interaction envelope organize visual action representations. Nat. Commun. 11, 1–11. https://doi.org/10.1038/s41467-020-16846-w

Thornton, M.A., Tamir, D.I., 2021. People accurately predict the transition probabilities between actions. Sci. Adv. 7, eabd4995. https://doi.org/10.1126/sciadv.abd4995

Thornton, M.A., Weaverdyck, M.E., Tamir, D.I., 2019. The brain represents people as the mental states they habitually experience. Nat. Commun. 10, 1–10. https://doi.org/10.1038/s41467-019-10309-7

Tucciarelli, R., Turella, L., Oosterhof, N.N., Weisz, N., Lingnau, A., 2015. MEG Multivariate Analysis Reveals Early Abstract Action Representations in the Lateral Occipitotemporal Cortex. J. Neurosci. 35, 16034–16045. https://doi.org/10.1523/JNEUROSCI.1422-15.2015

Tucciarelli, R., Wurm, M., Baccolo, E., Lingnau, A., 2019. The representational space of observed actions. Elife 8, 1–24. https://doi.org/10.7554/eLife.47686

Vangeneugden, J., Peelen, M. V., Tadin, D., Battelli, L., 2014. Distinct neural mechanisms for body form and body motion discriminations. J. Neurosci. 34, 574–585. https://doi.org/10.1523/JNEUROSCI.4032-13.2014

Wamain, Y., Pluciennicka, E., Kalénine, S., 2014. Temporal dynamics of action perception: Differences on ERP evoked by object-related and non-object-related actions.

Neuropsychologia 63, 249–258. https://doi.org/10.1016/j.neuropsychologia.2014.08.034

Weaverdyck, M.E., Thornton, M.A., Tamir, D.I., 2021. The representational structure of mental states generalizes across target people and stimulus modalities. Neuroimage 238, 118258. https://doi.org/10.1016/j.neuroimage.2021.118258

Willems, R.M., Peelen, M. V., 2021. How context changes the neural basis of perception and language. iScience 24, 102392. https://doi.org/10.1016/j.isci.2021.102392

Wurm, M.F., Caramazza, A., 2019. Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. Neuroimage 202. https://doi.org/10.1016/j.neuroimage.2019.116153

Wurm, M.F., Caramazza, A., Lingnau, A., 2017. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. J. Neurosci. 37, 562–575. https://doi.org/10.1523/JNEUROSCI.1717-16.2016

Wurm, M.F., Cramon, D.Y., Schubotz, R.I., 2012. The Context-Object-Manipulation triad: Cross Talk during action perception revealed by fMRI. J. Cogn. Neurosci. 24, 1548–1559. https://doi.org/10.1162/jocn_a_00232

Wurm, X.M.F., Lingnau, A., 2015. Decoding Actions at Different Levels of Abstraction. J. Neurosci. 35, 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15.