NeuroImage 243 (2021) 118534

Contents lists available at ScienceDirect

NeuroImage

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

Touching events predict human action segmentation in brain and behavior

Jennifer Pomp[a,b,*], Nina Heins[a,b], Ima Trempler[a,b], Tomas Kulvicius[c,d], Minija Tamosiunaite[c,e], Falko Mecklenbrauck[a], Moritz F. Wurm[f], Florentin Wörgötter[c], Ricarda I. Schubotz[a,b]

[a] *Department of Psychology, University of Münster, Germany*
[b] *Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Germany*
[c] *Institute for Physics 3 – Biophysics and Bernstein Center for Computational Neuroscience (BCCN), University of Göttingen, Germany*
[d] *University Medical Center Göttingen, Child and Adolescent Psychiatry and Psychotherapy, Göttingen, Germany*
[e] *Department of Informatics, Vytautas Magnus University, Kaunas, Lithuania*
[f] *Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Italy*

ARTICLE INFO

Keywords:
Action observation
Event segmentation
Unit marking procedure
fMRI
Semantic event chain
Computer vision

ABSTRACT

Recognizing the actions of others depends on segmentation into meaningful events. After decades of research in this area, it remains still unclear how humans do this and which brain areas support underlying processes. Here we show that a computer vision-based model of touching and untouching events can predict human behavior in segmenting object manipulation actions with high accuracy. Using this computational model and functional Magnetic Resonance Imaging (fMRI), we pinpoint the neural networks underlying this segmentation behavior during an implicit action observation task. Segmentation was announced by a strong increase of visual activity at touching events followed by the engagement of frontal, hippocampal and insula regions, signaling updating expectation at subsequent untouching events. Brain activity and behavior show that touching-untouching motifs are critical features for identifying the key elements of actions including object manipulations.

1. Introduction

Actions performed by others provide us with a continuous stream of complex perceptual input. Still, this stimulus entails a smoothly joined sequence of segments, which we can easily distinguish. Action observers expose an intra-individually highly consistent segmentation behavior when asked to indicate action steps by button presses (*unit marking procedure;* Newtson, 1973), suggesting that they perceive actions in stable units separated by breakpoints. These action segments have the tendency to preserve their integrity for instance by resisting interruptions (Newtson and Engquist, 1976) and missing content (Kosie and Baldwin, 2019), and being robust to perspective shifts (Swallow et al., 2018). Breakpoints systematically receive increased attention (Hard et al., 2011) and recognition memory for breakpoints is superior to that for other intervals (Swallow et al., 2009), probably because episodic memories emerge from significant contextual changes (Clewett and Davachi, 2017). This suggests that breakpoints contain more of the information from the continuous sequence than non-breakpoints and lead to the formation of new memory traces (Gershman et al., 2014). Moreover, breakpoints indicate that a distinctive change has occurred, rather than a distinctive state has been achieved (meaningful changes vs. meaningful states; Newtson et al., 1977). Event segmentation, applicable not only to observed actions but also to speech (Aslin, 2017; Wu and Bulut, 2020) or music (Sridharan et al., 2007), is suggested to efficiently improve predictions about the near future by integrating information over the recent past (Kurby and Zacks, 2008), and indeed, evidence of predictive action observation is abundant (e.g. Botvinick and Plaut, 2004; Colder, 2011; Csibra and Gergely, 2007; Graf et al., 2007; Kilner et al., 2007, 2004; Schiffer et al., 2013b; Stadler et al., 2011).

But what exactly determines how to segment an action into meaningful chunks? Humans spontaneously learn and use statistical information (Fiser et al., 2010; Perruchet and Pacton, 2006; Tobia et al., 2012), including 1st and 2nd level statistical structure during action observation (Ahlheim et al., 2014). A large repertoire of natural action segments could emerge simply from repeated experience of these segments in different contexts (Avrahami and Kareev, 1994). Importantly, breakpoints between action segments entail the most invariant stages of an action that occur in each effective action sequence (Byrne and Russon, 1998). Thus, breakpoints are reliable anchors in actions, but at the same time they mark the transition into phases of highest uncertainty, because different subsequent segments can be linked to the end of an action segment. Because the predictability regarding the further course of action

[*] Corresponding author.
E-mail addresses: jennifer.pomp@uni-muenster.de (J. Pomp), ima.trempler@googlemail.com (I. Trempler), tomas.kulvicius@phys.uni-goettingen.de (T. Kulvicius), minija.tamosiunaite@vdu.lt (M. Tamosiunaite), f_meck01@uni-muenster.de (F. Mecklenbrauck), moritz.wurm@unitn.it (M.F. Wurm), worgott@gwdg.de (F. Wörgötter), rschubotz@uni-muenster.de (R.I. Schubotz).

https://doi.org/10.1016/j.neuroimage.2021.118534.
Received 26 March 2021; Received in revised form 19 August 2021; Accepted 28 August 2021
Available online 29 August 2021.

is lowest at breakpoints, updating processes of the internal event model are presumably triggered exactly at this point in preparation for the coming action step (Kurby and Zacks, 2008; Schubotz et al., 2012). According to a recent model, event segmentation is driven by changes in inferences about what has generated them (Shin and DuBrow, 2021), making volatility, i.e., the inferred rate of change of the environment, a decisive factor regarding event segmentation (Hohwy et al., 2021). Breakpoints hence seem to be “stop and see” moments, where the completed action segment connects to the upcoming segment, and typically, there are several candidates for this upcoming segment, each having a certain probability.

Corroborating this assumption, it was found that brain activity during action observation varies as a function of the statistical structure provided by action segments (Schubotz et al., 2012). More specifically, the BOLD response increase reflects the level of quantified surprise at each breakpoint (Ahlheim et al., 2016; Ahlheim et al., 2014; Schiffer et al., 2013b, 2013), which has also been found in other paradigms as naturalistic movie perception (Brandman et al., 2021) and sports viewing (Antony et al., 2020). However, a crucial remaining question is exactly what kind of information drives human event segmentation. Functional MRI research suggests that *changes in motion* may serve as a core marker of breakpoints in actions, since brain areas specialized for motion processing, especially human motion area hMT, are significantly activated at breakpoints (Schubotz et al., 2012; Speer et al., 2003).

In the present fMRI study, we used a computer vision approach to directly test the assumption that human event segmentation relies on, and hence is predicted by, dynamic changes of the spatial relations between objects, hands and ground. Computer vision provides a unique avenue to objectively determine dynamic stimulus properties by extracting so-called *touching* and *untouching events* between objects (TUs, hereafter). Based on earlier works, our present approach provides a generic encoding scheme for object manipulations by constructing a dynamic graph sequence from continuously tracked RGB-D sensor data of action videos (Aksoy et al., 2011; Wörgötter et al., 2013). Topological transitions of these graphs occur whenever objects touch or untouch and are stored in a transition matrix called the *semantic event chain* (SEC). Crucially, this account is model-free and strictly stimulus-driven: It does not differentiate between hands, objects, or ground, nor does it require any functional or semantic knowledge about objects.

In a first step, a set of 48 object manipulations was recorded and subjected to a stimulus-driven segmentation of SEC events based on the extraction of TUs. In a second step, we presented 31 participants with the same videos in an fMRI study while they performed a cover task keeping their attention on the observed action. Subsequently, we conducted a test-retest procedure where the same group of participants engaged in a unit marking task, i.e. they indicated breakpoints in the action videos by button presses. We extracted those unit marks (Ms) that were consistently reported on group level (see Section 2.5.3 Determination of group-consistent unit marks for details). Finally, brain activity measured via fMRI was analyzed with regard to TUs and Ms. Using this approach, we aimed to determine to what degree brain activity and segmentation behavior in humans were linked to the event structure derived from computer vision.

We reasoned that if TUs are critical time points for action segmentation, then they should show a systematic relationship to Ms or even account for human segmentation behavior. Such a systematic relationship could mean that TUs and Ms temporally coincide or that we find a systematic temporal delay between both types of events. In case of coincidence, we expected to replicate previously found brain activation patterns for behaviorally determined action breakpoints, including increased engagement of motion sensitive area hMT, and in addition, also angular gyrus, superior frontal sulcus (SFS), and parahippocampal gyrus (PHG). While area hMT was found to increase at breakpoints also in coherent human motion in the form of Tai Chi videos, this fronto-parieto-hippocampal network became specifically engaged for breakpoints in goal-directed actions, presumably reflecting recall from semantic action knowledge (Schubotz et al., 2012). In the case that Ms and TUs do not or do not always coincide in time, we expected brain responses to differentiate between either type of event, allowing to dissociate the neural processes associated with TU analysis and segmentation decisions.

2. Methods

2.1. Participants

Thirty-one participants (M_{age} = 23.84 years, *SD* = 3.01, age range = 18 - 31 years, 25 women, 6 men) participated in the present study. The data of one additional participant was excluded from the analyses due to misunderstood instructions. All participants were right-handed as determined by the Edinburgh Handedness Inventory (Oldfield, 1971), had normal or corrected-to-normal vision, intact color perception, had no history of neurological or psychiatric diseases and met the criteria for MRI scanning. Twenty-nine of the participants were students. The local ethics committee of the Faculty of Psychology (University of Muenster, Germany) approved that the current study followed the principles set by the Declaration of Helsinki. The participants provided informed consent and either received course credits or were paid for their participation.

2.2. Stimulus material

The manipulation actions for the video stimuli were chosen according to the SEC framework (Wörgötter et al., 2013). Twelve actions were selected belonging to six action categories (see Supplementary Table 1 for a list of the individual object manipulations). Each action was recorded using four different objects which resulted in 48 object manipulations. Action videos were recorded using an industrial camera (BASLER acA 1300–75 gc) with a TV zoom lens (11.5 – 69 mm, 1:1.4) as well as an ASUS Xtion Live RGB-D sensor (ASUS TeK Computer Inc., Taipeh, Taiwan) recording color as well as depth images. For the video stimuli, the BASLER recordings were used, showing the actress from the front up to the shoulders performing the action on a white table. The ASUS Xtion Live recorded the actions from above and its recordings were utilized for TU time point extraction (see Section 2.3 Video Segmentation and SEC Determination). For each object manipulation six to seven unique video takes were chosen for the final stimulus set meaning that no video was repeatedly presented. In total, 294 action videos were shown to the participants. The videos had a frame rate of 23 fps. Each video started 10 frames before the hand lifts from the table to act and finished 5 frames after the hand lies back on the table with a video duration ranging from 72 frames to 185 frames (*M* = 114.79, *SD* = 19.74), i.e. 3130 ms to 8044 ms (*M* = 4991, *SD* = 858). To increase perceptual variability, the videos were mirrored so that actions seemed to be performed by the left hand. Each participant saw half of the actions mirrored.

The stimulus sequence was designed as a second-level counterbalanced De Bruijn sequence with seven conditions (6 action categories + null condition). Using the De Bruijn cycle generator by Aguirre and co-workers (Aguirre et al., 2011), 500,000 sequences were generated using NeuroDebian 8.0.0 (Halchenko and Hanke, 2012) and then the starting point of each sequence was shifted 47 times (length of the first run) resulting in 24,000,000 possible sequences of which the optimal one was chosen using a custom-built MATLAB R2019a (The MathWorks Inc., Natick, MA, USA) script. Subsequently, condition labels of the six experimental conditions were permuted to create 20 different stimulus lists. Per list, half of the stimuli were mirrored and a second list contained the complement of these which gave 40 different stimulus lists in total. For the second and third experimental session, the start of the individual stimulus sequence was shifted by one third and two third, respectively, to prevent recognition of the stimulus sequence as well as time-dependent effects. For the fMRI session, the stimulus sequence was subdivided into seven runs and at the start of each run the

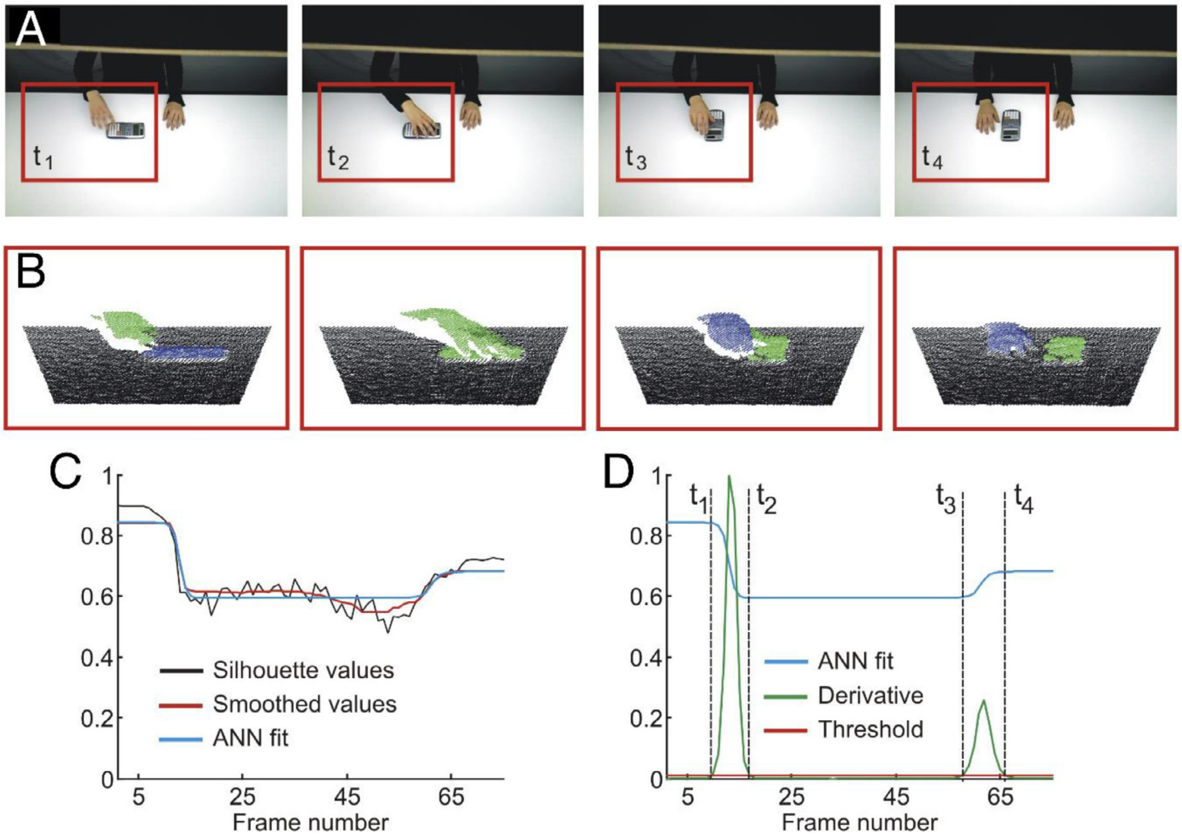

Fig. 1. Schema of the procedure for extracting the time points for touching and untouching events from an exemplary action, here "turning calculator". A) Point cloud extraction and preprocessing of RGB images. B) Clustering point clouds and calculating silhouette values. C) Curve fitting using artificial neural network (ANN): Raw silhouette values (black), smoothed silhouette values using median filter (red) and fitted silhouette curve using ANN (blue). D) Extraction of time events: Derivative of the ANN fit (green) and obtained time points of TU events after thresholding: t1 – hand detaches from the table (i.e., first untouching), t2 – hand touches calculator (i.e., first touching), t3 – hand detaches from the calculator (i.e., second untouching), and t4 – hand touches the table (i.e., second touching). Thus, in this example a U-T-U-T sequence is extracted. A demo source code of automated extraction that corresponds to the shown example can be downloaded from the OSF repository (accession code: https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9).

last two videos of the preceding run were repeated and then discarded from analyses to presume a continuous stimulus sequence.

2.3. *Video segmentation and SEC determination*

We used an automated extraction of time points of TU events, enabling a fast and accurate segmentation of action sequences based on objective criteria. A schema for the automated extraction of time points at which touching/untouching relations between object pairs change is shown in Fig. 1 and a demo source code underlying the example in Fig. 1 can be downloaded from the OSF repository (accession code: https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9). Here we used the frame number to define the time points. The input to the algorithm is a sequence of RGB-D frames *fi* (i = 1…n, n is the number of frames) and the output is a sequence of time events *ti* (i = 1…m, m is the number of TU events which was predefined manually). In the following subsections we provide details for the four main steps of the algorithm.

2.3.1. *Point cloud extraction and preprocessing*

Point clouds for each frame *fi* were generated from depth images which were acquired using ASUS Xtion Live sensor. Region of interest on the left side of the frame was cut as shown in Fig. 1, since always only one hand was involved in the analyzed actions. Furthermore, point clouds were subsampled by a factor of four in order to reduce the amount of points this way speeding up the clustering procedure. Before clustering, plane subtraction was performed. In most of the cases, ground plane subtraction (i.e., points corresponding to the table) was done by fitting flat 2D surface and then removing all points from the 3D point cloud data which were below the fitted ground plane (see black points in Fig. 1B). To be more specific, we removed points $p_i = \{x_i,y_i,z_i\}$, if z_i - $Z_i < th$, were $Z_i = P(x_i,y_i)$ are corresponding points of the fitted plane P, and $th=0.015$ is the ground plane threshold. The removed points p_i were not included to further cluster analysis. In some cases where very flat objects were present in the scene (e.g. a newspaper, playing card, etc.), we used color-based ground plane subtraction instead of the plane fitting procedure. Thus, for the clustering step, we only used point clouds of the hand and objects.

2.3.2. *Clustering and calculation of Silhouette scores*

Clustering of points (objects) was performed based on 3D point coordinates $p_i = \{x_i,y_i,z_i\}$ by using hierarchical clustering with Euclidean distance as a similarity measure and Ward's method as a linkage method. The clustering procedure was repeated K-1 times for each frame f_i (i = 1…n) with a predefined number of clusters k = 2…K, where K is the number of objects including the hand (but excluding the table). For each frame f_i we computed an average Silhouette score as follows:

$$S\left(f_i\right) = sum\left(S_k\right)/(K-1), \text{with} \tag{1}$$

$$S_k(j) = sum[(min(D_{between}(j,l)) - D_{within}(j))/\max(D_{within}(j), \min(D_{between}(j,l)))]/N, \tag{2}$$

where $D_{within}(j)$ is the average distance from the j-th point to the other points in its own cluster, and $D_{between}(j,l)$ is the average distance from the *j-th* point to points in another cluster *l*. Here *N* is the total number of points. The Silhouette score for each point *j* measures how similar that point is to points in its own cluster in comparison to points in other clusters. The values of the Silhouette score are between −1 and 1. Thus, when two clusters are getting closer, then the average score $S(f_i)$ decreases, while it increases when clusters are getting apart (see Fig. 1C). In this way, we used Silhouette values to find TU events. Note that the average silhouette value was less susceptible to noise in the point cloud data than the maximum value, resulting in a more accurate estimate of TU events. See the OSF repository (accession code: https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9) for a simulation of the differences between mean and maximum silhouette scores.

2.3.3. Fitting of Silhouette curve using ANN

The time points of TU events can be extracted from the Silhouette curve; however, Silhouette scores are noisy due to noise present in the point cloud data obtained from the RGB-D sensor. Thus, we first filtered the Silhouette scores $S(f_i)$ using a median filter with a time window of 20 frames and then fitted filtered scores with an artificial neural network (ANN). This leads to a smooth curve with descending and raising slopes which allows extracting of time points in the next step. For fitting $S(f_i)$, we used a fully connected feed-forward network with one hidden layer where in the hidden layer we used a *tansig* transfer function and in the output layer a *linear* transfer function was used. The number of neurons in the hidden layer corresponded to the number of sigmoid functions needed to fit the Silhouette value function S (see Fig. 1C,D), which corresponded to changes in cluster configuration, i.e., if two clusters are merging then objects are touching each other (T) and if two clusters are getting apart then objects are detaching from each other (U). In the given example in Fig. 1 for a “turn calculator” action, we have four TU events (hand lifts up from the table, hand touches the calculator, hand leaves the calculator, and hand touches the table). Thus, the TU events follow an irregular pattern of Ts and Us, and to represent two TU events one sigmoid function is needed as demonstrated by an example shown in Fig. 1D (see t1, t2 and t3, t4). The number of neurons h in the hidden layer was set based on the number of TU events m, i.e., h = round(m/2). In this case we used two neurons in the hidden layer. The network was fitted ten times and then the best outcome with respect to the minimal mean squared error between $S(f_i)$ and network's prediction $S_{ANN}(f_i)$ was used for the next step.

2.3.4. Extraction of time points

Finally, time points of TU events were extracted by applying dynamic thresholding to the derivative of the $S_{ANN}(f_i)$. We start with some initial threshold value TH_{ini} = 0.01 and increase it by 0.005 until the predefined number of TU time points is obtained. The time points are extracted at the frame numbers where derivative of the $S_{ANN}(f_i)$ crosses the threshold value *TH* (see Fig. 1D). The extracted time points were checked against manual segmentation results and time points whenever the algorithm misinterpreted the scene which gave an error message. Deviation from human segmentation on average was 3.49 frames (SD = 3.39), and in 94.45% of the cases deviation was less than ten frames (i.e., mean value + 2*SD). Thus, we corrected outliers in 5.55% of the cases, where event segmentation differences were larger than 9 frames by setting values of automated segmentation to corresponding values of human segmentation. The framework was implemented using MATLAB where standard MATLAB functions for clustering and ANN fitting were used. Extracted TU events were taken as machine-determined objective events (TUs) and the middle frames between two TU events were taken as non-events (nTU) to be maximally far away from an event.

2.4. Experimental procedure

Participants completed three sessions. The MRI session was on average 4 days (range = 3 - 7) before the behavioral test-retest sessions which were on average 14 days apart from one another (range = 14 - 17). During the first session, participants saw the action videos while being in the MR scanner. Action videos were back-projected onto a screen and presented centrally with a screen resolution of 640 × 512 pixels. Participants viewed the screen binocularly through a mirror above the head coil. Attention capturing questions regularly followed the videos asking whether an action description is appropriate for the just seen action video. Participants responded by pressing one of two response keys with their right index and middle finger. See Fig. 2A for the experimental trial design. Including anatomical scans and six short breaks during the task, the scanning time amounted to approximately 60 min. The overall duration of the first session was between 90 - 120 min including consent forms, instructions, preparation, scanning and a short survey at the end.

The second session comprised the unit marking task (Newtson, 1973). Participants saw the same videos as in the first session. Stimuli were presented on a 23″ monitor by Presentation 18.1 (Neurobehavioral Systems Inc., Berkley, CA, USA) and participants were instructed to press a button with their right index finger whenever they think an action step is finished, that is, a breakpoint occurred (cf. Schubotz et al., 2012). Training trials were offered at the beginning and two breaks were provided after one respectively two thirds of the trials. This task took approximately 45 min. See Fig. 2B for the experimental trial design. In the third session, this task was repeated to retest the unit marking behavior.

2.5. Behavioral data analysis

2.5.1. Intra-individual retest reliability of unit marking responses

The unit marking procedure is a subjective judgment task, so responses cannot be right or wrong. Therefore, retest reliability was assessed on single subject as well as on group level to ensure that responses were consistent and meaningful. In a first step, responses were converted from milliseconds to frames (one frame amounting to a 1000/23 ms segment) to allocate each button press to the correspondingly presented frame of the video. We did not subtract a hypothetical motor response time as participants were highly familiar with the kind of simple everyday actions that we employed, and this familiarity was even stronger in the behavioral sessions when participants saw the videos for the second respectively third time. Hence, we adopted the premise that responses were delivered in anticipation of critical events in the videos, not in a reactive manner.

On single subject level, we examined whether test responses matched retest responses consistently. To this end, trials in which the number of responses in the test session equaled the number of responses in the retest session were used to define an individual consistency criterion c_i, which was then applied to all trials independent of the number of responses. For each response in each of these same-number-of-responses-trials, the absolute difference $d_{|t-t'|}$ in frames between test button press *t* and retest button press *t'* was determined, and then averaged over all responses per participant. The upper bound of the 95% confidence interval (CI) of this mean difference score per participant was taken as individual criterion c_i for *consistent* button presses in the test and retest session. Thus, the individual criteria considered the individual variability in reaction times. To prevent too large cut-off values, we additionally calculated a global criterion c_g by averaging the individual criteria of our participants. The upper bound of the 95% CI of this average was used as global criterion c_g to threshold the individual consistency criteria c_i. If, for example, the individual criterion c_i of a participant was 14.5 frames but the global criterion c_g was 12.4 frames, the global criterion was applied for this participant. In sum, for each retest response *t'*, it was determined whether a test response t appeared within the individual time window around the retest response ($t' \pm c$). If this was the case, it was considered a consistent unit marking response. Subsequently, as a measure of single subject retest reliability, the percentage of consistent responses per participant was identified.

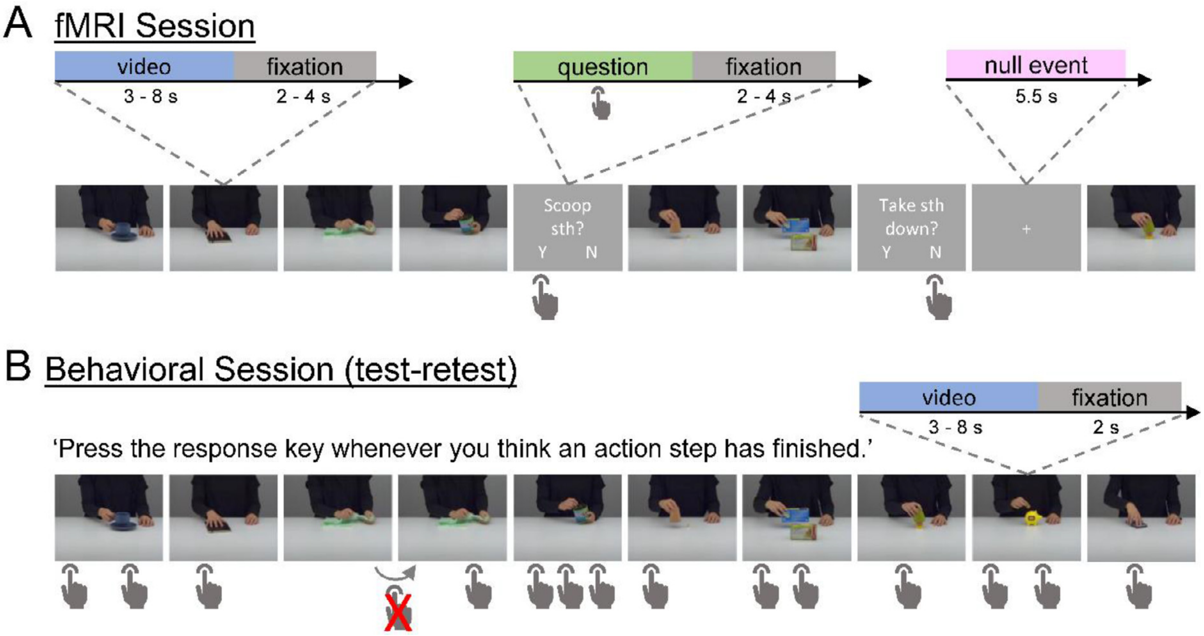

Fig. 2. Experimental design. (A) In the fMRI session, action trials and null trials were passively observed and question trials required participants to confirm or reject an action description with regard to the preceding action video. The question disappeared after button press. (B) In the two behavioral sessions (test-retest), participants saw the same videos as during fMRI and indicated by button press when they thought an action step had finished. In case no response was given, the video was repeated. Example videos are provided in an OSF repository (see https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9). The entire stimulus material is available via the Action Video Corpus Muenster (AVICOM, https://www.uni-muenster.de/IVV5PSY/AvicomSrv/).

To compare these results with random button presses, we in a first step shuffled the button press intervals. To this end, we extracted the time intervals between button presses (for the first button press in a video, we used the interval between this response and the video onset) in the test session per participant. From this distribution, we randomly drew and cumulated intervals to simulate random test session button presses while preserving the stochastic characteristics of the behavior. Using this procedure, we generated ten simulated test session data sets, calculated the percentage of consistent responses per participant (just like we did for the actual behavior) and averaged this percentage per participant over the ten simulations. To test whether participants performed more reliably than randomly, we calculated a paired-sample *t*-test between the actual percentage of consistent responses per participant and the percentages based on the simulated data.

2.5.2. Retest reliability of unit marking responses at the group level

To examine the unit marking responses at the group level, we smoothed the frame-by-frame data with a rectangular kernel with a width of three frames ($3^*(1000/23) \approx 130.4$ ms, referred to as *bin* hereafter). This means, for each video we aggregated the number of responses for each frame f_t plus those from adjacent frames f_{t-1} and f_{t+1}. Thereby we pooled the data of all participants. A maximum of one response per participant was included in a bin of three frames, so that the maximum value a bin could reach was equal to the total number of participants ($n = 31$). The bin value was then allocated to the middle frame f_t of the bin and will be referred to as *frame value* hereafter. Consequently, the frame value was set to zero if no response had occurred within the bin.

To determine the group level retest reliability, we correlated the time series of frame values per video between the test and the retest sessions (Pearson's *r*). The *r*-values per video were then Fisher *z*-transformed, averaged and retransformed to *r* to give a mean correlation.

2.5.3. Determination of group-consistent unit marks

The maximum frame value of an action video was taken to indicate group-consistent unit marks (M). Fig. 3 shows the time series of frame values based on individual unit markings for two example videos with corresponding group-consistent Ms at maximum frame values as well as objective TU events to illustrate their temporal distribution. In order to objectify the maximum frame values, we utilized the ten simulated test session data sets that were generated to evaluate single subject retest reliability (cf. Section 2.5.1 Intra-individual retest reliability of unit marking responses). We applied the same protocol to these ten simulated data sets as we did to the original data to determine group-consistent unit marks and compared the resulting maximum frame values to the actual ones. To determine the non-unit-mark (nM) for the fMRI analyses, one of the frames with the minimum frame value of zero was randomly chosen excluding the first 12 and last 12 frames of each video. Ms and nMs were then used to model brain responses.

2.5.4. Convergence of human-determined unit marks (M) and objective events (TU)

The hypothesis of dependence of human action segmentation (M) on objective touching and untouching events (TU) was tested by analyzing the relationship between human-determined unit markers and objective events in several steps. To evaluate whether the majority of Ms coincides with TUs, we examined how often a TU was not further than two frames (i.e. maximally ~130 ms) away from an M. Subsequently, we compared this result to randomly distributed unit marks. As with the test-retest performance of individual subjects, we shuffled the time intervals generated by the unit marks and randomly drew from this shuffled distribution to simulate random unit marks while preserving the stochastic characteristics of the group behavior. We generated ten simulated data sets containing unit marks, examined individually how often a TU was no more than two frames away from a simulated M, and then calculated a one-sample *t*-test to compare the resulting coincidence rates with the coincidence rate of the actual unit mark distribution. In addition, we examined whether the TU closest to an M in each case precedes ("pre-M") or follows ("post-M") this M, provided that the M and TU events did not fall at exactly the same time.

Based on the outcome of this analysis (as described in the Results section), we examined the temporal relationship between M and TU events

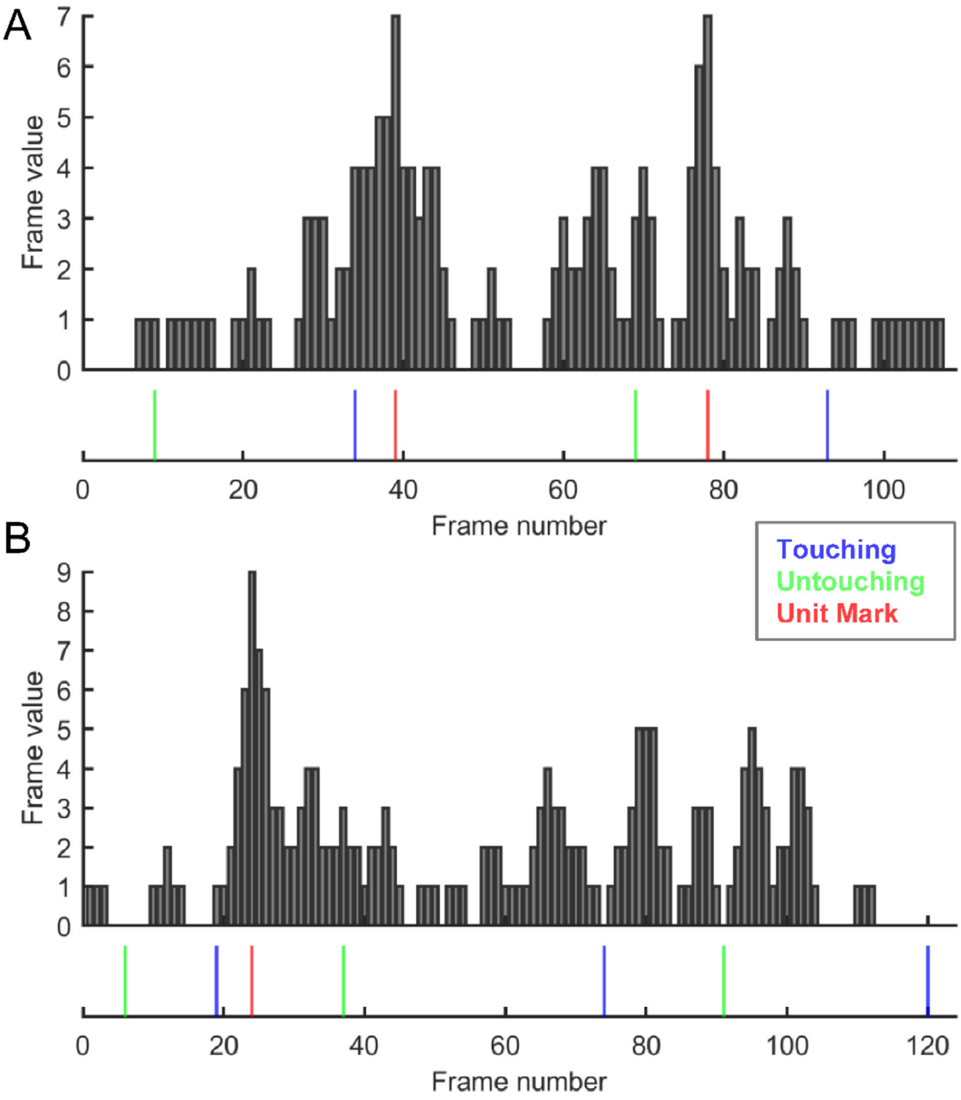

Fig. 3. Pooled unit marking responses of the group (n = 31) for two exemplary object manipulation videos: turning a bottle (A), putting a cup on top of a saucer (B). Maximum frame values were taken as group-consistent unit marks (Ms), as indicated in red on the lower x-axis. Respective touching (T) and untouching (U) events are given in blue and green.

in more detail in the following way. Firstly, for the closest TU of each M, we determined: (a) the direction of time lag (pre-M; post-M), and (b) the type of TU (touching, T; untouching, U). Secondly, we determined the temporal distance between Ms and the TUs events preceding and following it. Thirdly, to test whether Ms have a systematic temporal relationship only to Ts but not Us, or vice versa, we determined separately for each M the temporally closest touching respectively untouching event and inspected their temporal distribution.

2.5.5. *Identification of sequential TU motifs embedding unit markings*

Finally, the same close-M touching and untouching events were examined with regard to typical sequential motifs embedding Ms using RStudio (Version 1.3.959, RStudio, PBC, Boston, MA) to identify stimulus-based (objective) reasons for reporting an event boundary. We introduce the term "motif" for a sequence of T and U events that embed M events more than randomly often. For this purpose, the two TUs preceding an M and one TU following an M were taken into account yielding a TU-TU-M-TU event scheme (e.g., T-U-M-T, T-T-M-U or U-T-M-U). This event scheme was chosen for several reasons. First, M events were preceded by at least one and at most two events in most videos (see the plot in Fig. 5 and see also Table 2 in the Supplementary Material for a list of all possible triplets and their probability of embedding an M). We therefore included two TU events *before* Ms in the analysis. The event scheme was then analyzed to clarify whether the occurrence of Ms systematically depended on one or two preceding TU events, as formulated in the hypotheses. In addition, one TU event *after* M was considered in each case to distinguish whether Ms occurred only in response to TU events or whether they also indicated (predictively) the occurrence of an upcoming TU event.

Considering the general likelihood of occurrence of such TU-TU-TU sequential triplets, we now explored whether any of these triplets was more likely to embed an M than could be expected from its general (stochastic) likelihood. To this end, we performed a chi-square test using SPSS 26 (IBM, New York, USA) to determine whether the proportions of TU-TU-TU triplets embedding an M differ from the general likelihood of occurrence of these triplets. Subsequently, we ran post hoc chi-square tests on single cells adjusting the significance values by multiplying by the original number of cells to account for multiple comparisons. This analysis identified sequential motifs that significantly co-occured with Ms.

2.5.6. *Manual video content analysis of sequential triplets*

For descriptive reasons, we also examined the content of the most frequently occurring M-embedding motifs. Since object identity was relevant for this, this mapping had to be done manually, as the algorithm does not distinguish between objects. For this video content analysis, we first defined the phases of transport and manipulation as ‘hand transport’ (from untouching of the hand until it touches again), ‘object transport’ (from untouching of the object until it touches again), ‘object manipulation’ (from hand touching the object until it untouches after manipulation), and ‘tool transport’ (hand with tool untouches until tool touches object); then we defined the phases where the hand or the tool is in contact with the object without moving (transporting or manipulating) as ‘start of object transport’ (from hand touching object until object untouching to be transported), ‘end of object transport’ (from object touching at the end of transport until hand untouching the object), and ‘end of manipulation with a tool’ (from untouching of a part of the object to untouching of the tool and the object). For the Ms embedded in T-U-X sequences (i.e., sequences of three events which start with a T followed by a U and then X stands for either T, U or the end of the video), either in the first or in the second phase, we extracted the corresponding action phase and compared the occurrence rates with the general likelihood of occurrence of these phases using Pearson’s chi-squared test and post hoc chi-square tests on single cells adjusting the significance values by multiplying by the original number of cells to account for multiple comparisons.

2.6. *fMRI data analysis*

2.6.1. *fMRI data acquisition and preprocessing*

Functional MRI data were acquired using a 3-Tesla Siemens Magnetom Prisma MR tomograph (Siemens, Erlangen, Germany) with a 20-channel head coil. Prior to functional imaging, a 3D-multiplanar rapidly acquired gradient-echo (MPRAGE) sequence was run to obtain high resolution T1-weighted images (scanning parameters: 192 slices, TR = 2130 ms, TE = 2.28 ms, slice thickness = 1 mm, FoV = 256 × 256 mm^2, flip angle = 8°). Blood-oxygen-level-dependent (BOLD) contrast was measured by gradient-echo echoplanar imaging (EPI). Seven EPI sequences were used to measure the seven experimental blocks (scanning parameters: 33 slices, TR = 2000 ms, TE = 30 ms, slice thickness = 3 mm, FoV = 192 × 192 mm^2, flip angle = 90°).

Anatomical and functional images were preprocessed using the Statistical Parametric Mapping software (SPM12; The Wellcome centre for Human Neuroimaging, London, UK) implemented in MATLAB R2019a. Preprocessing included slice time correction to the first slice, realignment to the mean image, co-registration of the functional images to the individual structural scan, normalization into the standard anatomical MNI space (Montreal Neurological Institute, Montreal, QC, Canada) on the basis of segmentation parameters, as well as spatial smoothing using an isotropic 8 mm full-width at half maximum (FWHM) Gaussian kernel. To remove low-frequency noise, a 128 s temporal high-pass filter was applied to the time-series of functional images.

2.6.2. *fMRI design specification*

Statistical analyses of functional images were done using SPM12 implementing a general linear model (GLM) for serially autocorrelated observations (Friston et al., 1994; Worsley and Friston, 1995) and a convolution with the canonical hemodynamic response function (HRF). In each GLM, the six subject-specific rigid-body transformations obtained from realignment were utilized as regressors of no interest. The volumes of the first two video presentations of each EPI were discarded to allow for T1-equilibrium effects.

To investigate functional areas specialized in the processing of action boundaries, a GLM was constructed including eight regressors of interest coding for onsets and durations of the specific event types: video, group-consistent unit mark in the test-retest session (M), no unit mark in the test-retest session (nM), objective touching event (T), objective untouching event (U), non-TU (nTU), null event and question. For each of the 350 Ms, a nM was determined (n = 350) (see *Section 2.5.3* Determination of group-consistent unit marks) and included in the design. Likewise, all 814 touching and all 772 untouching events were included and correspondingly 772 nTUs (see Section 2.3 Video segmentation and SEC determination). Both types of non-critical events (nTU and nM) appeared distributed over the video duration (Supplementary Figure 1) and were chosen to be maximally far away from their corresponding events (i.e., as nTUs, the frame in the mid between two TU events were chosen and as nMs, frames where no participant marked a unit). Group-consistent unit marks instead of individual unit marking responses were chosen to model the data to obtain a more stable model.

To prevent basic and object motion as well as effects of the mere time point in the video from confounding our analyses, we considered several factors in the choice of non-critical events and benefitted from the natural structure of our events. First, hMT was among the regions we expected to show increased activity at action boundaries. Previous studies reported that activity in hMT increases at event-segment boundaries, suggesting that motion information is processed particularly intensively here (Schubotz et al., 2012; Speer et al., 2003; Zacks et al., 2006). However, to interpret the increased activity in hMT at action boundaries in this sense, it must be ruled out that this effect is merely due to an increase in motion in the stimulus. This can already be assumed theoretically, since TU events are accompanied by a sharp slowdown or even a complete stop of the movement. However, to show this empirically, we performed a dense optical flow analysis for each video and tested the correlation between the optical flow values and the binary vectors of touching events and untouching events (1 = T/U, 0 = nT/nU). We then calculated t tests on r-values across all videos. As a result, we found a weak but highly significant negative correlation of optical flow with touching events ($t(293) = -5.7$, $p < .001$, mean $r = -0.02$) and no significant correlation of optical flow with untouching events ($t(293) = -1.4$, $p = .174$, mean $r = -0.006$). In addition, we tested for the same correlation effects based on the concatenated vectors of all videos, which also revealed a weak but significant correlation of optical flow with concatenated touching events ($r(33{,}748) = -0.02$, $p < .001$) and no such effect for concatenated untouching events ($r(33748) = -0.005$, $p = .361$). Thus, as suspected, a weak but clearly significant negative correlation of motion and T events was found. Although such a weak correlation should be interpreted with caution, it allows us to rule out the possibility that T events were associated with an increase in motion in the stimulus.

Secondly, neither TU events nor M, nTU or nM events did systematically occur only at the beginning or the end of the videos, but were distributed across the entire video duration (Fig. 5, Supplementary Figure 1). Relative to the length of the video, the earliest M appeared after 19% of the video and the latest M at the end of the video (M = 50%, SD = 23). The earliest nM appeared after 11% and the latest after 90% (M = 45%, SD = 23). Analogously, the earliest TU event appeared after 2% and the latest at the end of the video (M = 50%, SD = 30) and the earliest nTU event appeared after 7% and the latest after 94% (M = 50%, SD = 25).

On the first level, t-contrasts for Ms versus nMs were calculated and submitted to a second-level t-test to detect functional areas specialized in the processing of action boundaries on group level. Analogously, t-contrasts for T versus nTU and U versus nTU were conducted. Furthermore, we contrasted all TUs (T + U) versus nTUs to detect areas specialized for both touching and untouching. To assure the specificity of these results, we calculated t-contrasts for the direct comparison between human-determined and objective events which means the conjunction of M versus T and M versus U (M>T ∩ M>U), the direct contrast of T versus M (T>M) and the direct contrast of U versus M (U>M).

Because the fMRI design described above considered only M events that occurred consistently across the whole group (cf. Section 2.5.3 Determination of group-consistent unit marks), one could argue that our analysis did not consider local peaks that could well indicate equally significant agreement between subjects. For this reason, we created another design as a control, an additional GLM including a regressor for video frame onset with a parametric modulator considering all individual unit marks Mp (*parametric unit mark)*. This parametric modulator indicated the continuous moment-by-moment fluctuation of unit marking responses of all subjects (number of unit marking responses relative to number of participants, e.g. 5/31, 2/31 and so forth) instead of binarized Ms and nMs, and replaced the regressors video, group-consistent unit mark in the test-retest session (M) and no unit mark in the test-retest session (nM). We then generated *t*-contrasts for Mp, as well as for the other contrasts of interest to control for the impact of modeling Ms parametrically, including T versus nTU, U versus nTU and TU versus nTU.

For all contrasts, we applied explicit gray matter masking on the first level. Therefore, we smoothed the individual normalized gray matter image at 8 mm FWHM and created a binary mask with a threshold of 0.2 using SPM12, as proposed by Jonathan Erik Peelle (http://jpeelle.net/mri/misc/creating_explicit_mask.html). For second-level whole-brain analyses, false discovery rate (FDR) correction at $p < .005$ peak level and a cluster extent threshold of 15 voxels was applied. Activity patterns were visualized using MRIcroGL 3D visualization software (McCausland Center for Brain Imaging, University of South Carolina, USA). Unthresholded statistical maps have been uploaded to NeuroVault.org (Gorgolewski et al., 2015) and are available at https://neurovault.org/collections/8736.

3. Results

3.1. Behavioral results

3.1.1. Intra-individual retest reliability of unit marking responses

Regarding single-subject level retest reliability, on average 62.99% were consistent responses (i.e., the test response matched the retest response in time) ranging between the participants from minimally 33.73% to maximally 87.56% (*SD* = 9.13). The individual consistency criterion c_i that defined the width of the time window around the retest response individually for each participant was minimum 4.6 frames (i.e., ~200 ms), median 8.5 frames (i.e., ~370 ms) and set to a global maximum c_g of 13 frames (i.e., ~565 ms), i.e., the rounded up upper bound of the 95% CI of the individual criteria (95% CI [7.98, 12.36]). Importantly, the consistency of the participants' unit marking behavior was significantly better than random button presses ($t(30) = 10.6$, 95%-CI [17.11,25.24], $p < .001$, $d = 1.91$, *two-sided)*. In sum, human unit marking was intra-individually consistent across the test-retest sessions.

3.1.2. Retest reliability of unit marking responses at the group level

Correspondingly, between-subjects unit marking behavior was consistent, as revealed by a significant correlation between group-based test-retest segmentation performance. Correlations testing the group level retest reliability yielded a mean correlation of test and retest smoothed time series of frame values per video of $r_z(292) = 0.55$ ($r_{\min} = 0.19$, $r_{\max} = 0.86$; each individual correlation per video being significant, all $p <= 0.04$).

3.1.3. Determination of group-consistent unit marks

The frame with the maximum frame value in a video that represents the maximum agreement between participants was taken as group-consistent M. On average this maximum frame value was 8.05 (*SD* = 1.82) ranging from 5 to 14. All maximum frame values were at least two standard deviations above the mean frame value of the respective video, which is in line with previous approaches (Schubotz et al., 2012). The maximum frame values resulting from simulated random unit markings ranged on average from 5.70 to 5.87 which was clearly below 8.04. In none of the simulated data sets were the maximum frame values two standard deviations above the respective video mean. This suggests that the subjects did not segment the videos randomly. The number of Ms per video on group level ranged from 1.0 to 4.0 with a mean of 1.2 (*SD* = 0.45, *n* = 294) and was significantly lower ($t(586) = 67.2$, 95%-CI [−4.33,−4.08], $p < .001$, $d = 5.55$, *two-sided*) than the number of TUs per video that ranged from three to seven (*M* = 5.4, *SD* = 0.97, *n* = 294). On single-subject level, the average number of individual test-retest consistent unit marking responses per video ranged from 0.7 to 1.8 with a mean of 1.3 (*SD* = 0.21, *n* = 294). Importantly, the number of individually consistent unit marking responses per action significantly correlated with the number of TUs per action ($r(292) = 0.52$, $p < .001$), pointing to a systematic relationship between the number of Ms and TUs.

3.1.4. Temporal relationship between Ms and TUs

With regard to the temporal relation between Ms and TUs, for about one third (28.3%) of the Ms, the time lag to the next TU was maximally two frames, i.e., up to ±130 ms. This coincidence rate was higher than the coincidence rate generated by random unit marks ($t(9) = -4.0$, 95%-CI [23.23,26.88], $p = .003$, $d = 1.27$, *two-sided*). Accordingly, Ms were systematically delivered in relation to TUs which was in line with our expectation.

Regarding the temporal relationship of Ms and their closest TUs on macroscopic level, we found that Ms followed TUs with a mean latency of 6.2 frames (*SD* = 4.5; i.e., 268 ± 195 ms) and preceded TUs with a mean latency of 4.5 frames (*SD* = 3.4; i.e., 196 ± 147 ms). Moreover, we found the majority (73%) of Ms to follow a TU; among these cases, there was a bias towards following a touching event (45%) vs. following an untouching event (28%). Ms that preceded the closest TU (22%) mostly did so for untouching events (17%) but rarely for touching (5%). The exact temporal distribution of pre-M and post-M objective events differentiated for touching and untouching revealed that if the closest TU to an M was a touching event, it mostly preceded the M (*Median* = −5 frames or ~217 ms). In cases where the closest TU to an M was an untouching event, its likelihood of occurrence peaked closer to the M (*Median* = −2 frames or ~87 ms). Furthermore, the dispersion for touching events (*SD* = 5.5) was descriptively smaller than for untouching events (*SD* = 6.0). Examining the likelihood of occurrence of close-M touching and close-M untouching events separately (Fig. 4), this pattern became even clearer. Close-M touching events more sharply preceded the M (*Median* = −6, *SD* = 13.3) whereas close-M untouching events more widely scattered around Ms with a slight precedence bias (*Median* = −2, *SD* = 17.3). These findings suggest that Ms often followed a T or scattered around a U event.

3.1.5. Sequential TU motifs typically embedding Ms

A major goal of our study was to identify stimulus-based (objective) reasons for reporting an event boundary. Thus, our approach was to examine the systematic relationship between touching and untouching on the one hand and Ms on the other. To test that this relationship was not random, we tested whether the frequency of an M-embedding TU scheme (TU-TU-M-TU) was significantly different from its purely stochastic occurrence probability (independent of its cooccurrence with an M) in the experiment. The analysis of the TU-TU-TU sequential triplets with regard to their embedding Ms revealed that of all possible TU-TU-M-TU event schemes, some were more likely to embed an M than others, and these were T-U-M-TU (i.e., first a T, then a U, then an M, and then either a T or a U) and TU-T-M-U (i.e., either a T or U at the beginning and then a T, an M and a U) sequences. Thus, most of the Ms (80%) coincided with a *touching-untouching (T-U) motif* (either T-U-M or T-M-U) within these triplets. This highlights the relevance of T-U motifs, where Ms occur either between T and U (T-M-U) or after T-U (T-U-M). Importantly, the proportion of triplets embedding an M significantly differed from the general likelihood of occurrence of these triplets ($\chi^2(6) = 67.03$, $p < .001$, Cramer's $V = 0.46$, $n = 314$). Post hoc single cell tests showed that the triplets U-T-U ($\chi^2(1) = 28.55$, $p < .001$, Cramer's $V = 0.30$, $n = 314$) and T-U-U ($\chi^2(1) = 12.32$, $p = .003$, Cramer's $V = 0.20$, $n = 314$) embedded Ms more frequently than expected and the triplet T-U-T ($\chi^2(1) = 38.17$, $p < .001$, Cramer's $V = 0.35$, $n = 314$) less frequently than expected, based on the general likelihood of occurrence of these triplets. See Supplementary Table 2 for the observed and expected numbers. Thirty-six Ms did not have two TU events before and one TU event after it such that they were not included in the

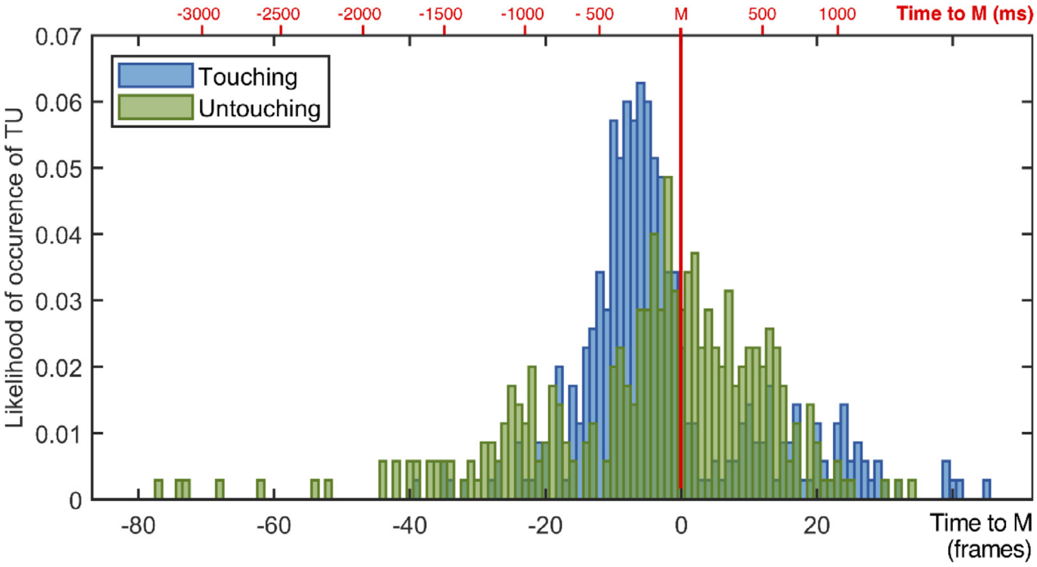

Fig. 4. Likelihood of occurrence of M-close touching events and M-close untouching events; the solid red line indicates the point in time where participants delivered a response for unit markings in the test-retest sessions (M), the lower x-axis shows the temporal distance of the events to M in frames and the upper x-axis additionally gives milliseconds for orientation.

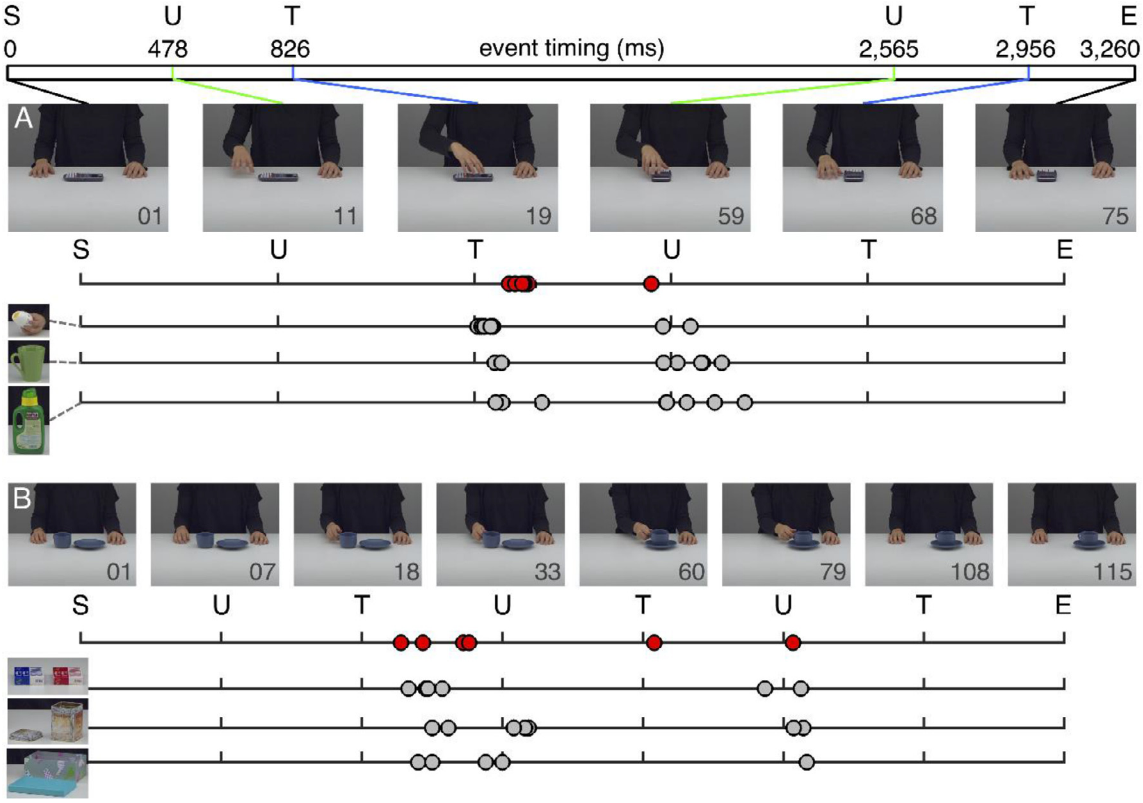

Fig. 5. Touching (T) and untouching (U) events as determined by computer vision for two exemplary object manipulation videos, and corresponding unit marks (M) delivered by participants. Single frame images are shown for all identified T and U events, with frame numbers given in the downright corner of the respective image. X-axes show Ms delivered relative to TU events (i.e., distances between TU events are warped and Ms are plotted according to their proportional timing between two events); *S* = Start, *U* = Untouching, *T* = Touching, *E* = End. A) "Turning calculator" action with Ms on the upper x-axis in red and Ms for the other three objects (i.e., an egg timer, a mug, a bottle) being turned on the lower three x-axes in gray. The horizontal bar above the single frame images shows the actual temporal distribution of the TU events across the action video in milliseconds as also given in the frame numbers (1 frame lasted approximately 43.5 ms). B) Correspondingly, "putting cup on top" action showing the Ms for the cup-using action on the upper x-axis in red and the Ms for the other three objects being put on top (i.e., two packs of playing cards, the lid of a tea tin, the lid of a container) on the lower three x-axes in gray.

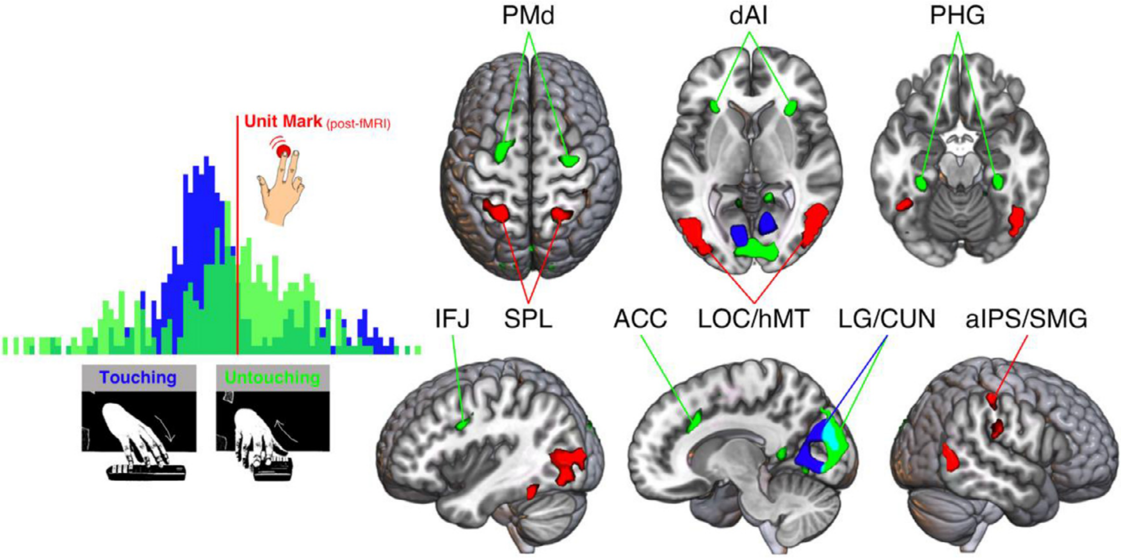

Fig. 6. Functional MRI activation at $p < .005$, peak-level FDR-corrected, for the main contrasts of post-fMRI human-determined unit marks (M>nM, red), objective touching events (T>nTU, blue) and objective untouching events (U>nTU, green). The overlap of the activation of touching and untouching in the LG/CUN region is shown additively in cyan. PMd = dorsal premotor cortex, dAI = dorsal anterior insula, PHG = parahippocampal gyrus, IFJ = inferior frontal junction, SPL = superior parietal lobule, LG = lingual gyrus, CUN = cuneus, LOC = lateral occipital cortex, hMT = motion area, ACC = anterior cingulate cortex, aIPS = anterior inferior parietal sulcus, SMG = supramarginal gyrus. Unthresholded statistical maps have been uploaded to NeuroVault.org and are available at https://neurovault.org/collections/8736.

analysis of sequential motifs. Fig. 5 shows the distribution of Ms relative to TU events exemplified by two object manipulations. Please note that the delay between events is displayed in a warped fashion and does not show the temporal distribution of the TU events in the course of the action video. In sum, our results showed that Ms coincided with a T-U motif disproportionately often, i.e., significantly more often than would have been expected based on their frequency of occurrence. We can conclude from these findings that people usually locate action boundaries exactly where a touching-untouching motif occurs in contrast to, for instance, untouching-untouching sequences.

3.1.6. *Action phases typically embedding Ms*

As 80% of the Ms appeared in either T-U-M or T-M-U, we had a closer look at T-U-X sequences (where X stands for either T, U or the end of the video) embedding an M either in the first or in the second phase. The respective video content analysis of the time between T-U and U-X revealed that the observed action phases embedding an M significantly differed from the general likelihood of occurrence of these action phases ($\chi^2(6) = 89.16$, $p < .001$, Cramer's $V = 0.57$, $n = 279$) (Supplementary Table 3). Post hoc single cell tests showed that Ms were more frequently than expected placed in phases of object manipulation ($\chi^2(1) = 34.72$, $p < .001$, Cramer's $V = 0.35$, $n = 279$) and at the start of object transport ($\chi^2(1) = 34.16$, $p < .001$, Cramer's $V = 0.35$, $n = 279$) while being less frequently than expected placed in phases of hand transport ($\chi^2(1) = 14.81$, $p < .001$, Cramer's $V = 0.23$, $n = 279$), object transport ($\chi^2(1) = 9.91$, $p = .012$, Cramer's $V = 0.19$, $n = 279$) and at the end of object transport ($\chi^2(1) = 13.60$, $p = .002$, Cramer's $V = 0.22$, $n = 279$). Overall, the only action phases in which subjects emitted significantly more Ms than statistically expected were during object manipulation and at the beginning of object transport.

Together, this pattern of results clearly shows a systematic temporal relationship between TUs and M. It suggests that participants pressed the button for action segmentation in response to sequential T-U motifs that indicate object manipulation or the start of object transport. Still, there were many more TUs than Ms, and consequently, the majority of TUs did not relate to an M. This allowed a clear dissociation of the neural processes associated with TU analysis and segmentation decisions.

3.2. *fMRI results*

In order to neither over- nor underestimate differences between T, U and M events, we considered each event in contrast to unspecific points in time between them (nTU and nM) as well as the conjunctions of direct contrasts for M (M>T ∩ M>U) and direct contrasts for T (T>M) and U (U>M). Hence, our discussion is restricted to brain activity uniquely observed for each of these three event classes.

To identify the network associated with unit marking in post-fMRI test-retesting, we ran a whole-brain analysis of the contrast M>nM (Fig. 6) which revealed significant bilateral activation in the lateral occipital cortex (LOC) comprising hMT (see e.g. Tootell et al., 1995, reporting similar peak coordinates; Table 2), the superior parietal lobule (SPL) and significant unilateral activation in the left fusiform gyrus (FG), right anterior inferior parietal sulcus (aIPS) and right supramarginal gyrus (SMG).

To address the brain response to the objective touching and untouching events, we calculated the contrast TU>nTU that yielded a bilateral activation cluster including the cuneus, lingual gyri and right parahippocampal gyrus. This cluster had no overlap with the pattern found for unit marks (M>nM).

Examining TU events in more detail, we separately computed T>nTU and U>nTU. The brain response to touching events (T>nTU; Fig. 6) showed a bilateral activity pattern in secondary visual areas spanning lingual gyri and cuneus. The brain response to untouching events (U>nTU; Fig. 6) showed a more extended network going beyond the cluster of lingual gyrus and cuneus also identified for T>nTU. This untouching specific activity comprised parahippocampal gyrus (PHG), the parieto-occipital fissure, dorsal premotor cortex (PMd), right anterior SFS (aSFS), left inferior frontal junction (IFJ), the right dorsal anterior cingulate cortex (dACC), and dorsal anterior insula (aAI). See Table 1 for the peak maxima of the described main contrasts.

The additionally calculated direct contrasts between human-determined and objective events validated the specificity of the above findings (Supplementary Figure 2). The conjunction of M>T ∩ M>U largely yielded the same pattern as M>nM with LOC/hMT, SPL, FG, aIPS/SMG, and furthermore found the ventral premotor cortex

Table 1
Maxima of activation from the main contrasts of the second-level whole-brain analyses at p <0.005 peak-level FDR-corrected.

Macroanatomical Location	H	Cluster Extent	*t*-value	MNI Coordinates x	y	z
M > nM						
Lateral occipital cortex / human motion area	L	335	9.30	−48	−73	−4
	R	452	9.25	51	−64	−7
Fusiform gyrus	L	40	6.63	−48	−52	−19
Superior parietal lobule	L	126	6.84	−24	−52	68
	R	102	7.02	18	−55	68
Anterior inferior parietal sulcus	R	27	5.24	54	−25	50
Supramarginal gyrus	R	44	4.74	57	−25	20
TU > nTU						
Cuneus	L	1491	8.56	−9	−97	17
	R		8.45	15	−94	29
Lingual gyrus	L		7.58	−6	−79	−1
	R		5.80	12	−79	−4
Parahippocampal gyrus	R		4.88	30	−37	−16
T > nTU						
Lingual gyrus	L	577	7.82	−9	−76	−1
	R		6.43	12	−76	−4
Cuneus	L		6.97	−9	−88	23
	R		6.50	9	−76	26
U > nTU						
Lingual gyrus	L	1522	9.82	−24	−73	−4
	R		8.90	33	−52	−7
Cuneus	L		8.74	−9	−100	17
	R		8.38	15	−94	29
Parieto-occipital fissure	L	68	5.15	−21	−58	14
Parahippocampal gyrus	L		6.23	−30	−34	−16
	R		5.66	30	−31	−16
Dorsal premotor cortex	L	204	7.39	−24	2	53
	R	174	6.54	24	2	50
Anterior superior frontal sulcus	R	20	5.07	27	35	29
Inferior frontal junction	L	27	5.22	−36	5	29
Dorsal anterior insula	L	31	6.03	−27	23	−1
	R	74	6.22	30	23	5
Dorsal anterior cingulate cortex	R	38	5.73	12	20	32

Note. H = Hemisphere, MNI = Montreal Neurological Institute, L = Left, R = Right, M = Unit mark, nM = non-unit mark, T = touching event, U = untouching event, nTU = non-touching/untouching event.

(PMv) / inferior frontal gyrus (IFG) and mid-insula to be activated. The direct contrast of T>M revealed the same pattern as T>nTU including bilateral lingual gyrus and cuneus. Finally, the direct contrast of U>M largely reflected the above referred findings for U>nTU yielding cuneus activation, the parieto-occipital fissure, PHG, PMd, aSFS, dAI, and ACC. See Supplementary Table 4 for the peak maxima of these direct contrasts.

The additionally calculated parametric GLM, considering all individual unit marking responses as a cumulative parametric regressor Mp, replicated and validated the specificity of the above findings. Investigating unit marks as parametric modulator based on the time series of the pooled unit marking responses revealed the same pattern as M>nM with LOC/hMT, FG, SPL, SMG, and furthermore yielded additional activity in angular gyrus, dorsal premotor cortex, and left IFG. All other contrasts (TU>nTU, T>nTU, and U>nTU) remained unchanged (see Supplementary Table 5 for the peak maxima of all contrasts from this GLM).

To summarize the fMRI results, we found distinct activity patterns for touching and untouching events which both clearly deviated from the network activated by the (independently tested) unit mark processing. Touching events' activity pattern comprised secondary visual activation and untouching events' activity pattern extended this network to parahippocampal, dorsal prefrontal, medial frontal and insular regions. In contrast, unit marks (as determined in the post-fMRI test-retest sessions) revealed increased activity of LOC, FG and parietal regions. The direct contrasts between Ms, Ts and Us corroborated differentiability of these events.

4. Discussion

The present study used computer vision methods to investigate whether human action segmentation behavior can be traced to objectifiable events of touching and untouching and fMRI to investigate the neural basis for processing these events. Participants watched videos of object-directed actions in an fMRI session, and subsequently two more times in a behavioral test-retest regime to ensure reliability of the determined Ms and to model brain activity at M. In the same set of action videos, the occurrences of touching and untouching events were determined based on a computer vision algorithm. Our results indicate that touching-untouching motifs can predict human action segmentation and are processed in distinct networks. Both behavioral effects as well as BOLD responses were highly informative with regard to the question whether touching and untouching events can help to objectify human action segmentation, as will be discussed in the following.

Considering first the behavioral results, the test-retest procedure following the fMRI session revealed that humans' action segmentations were relatively consistent both on the individual as on the group level (cf. Schubotz et al., 2012). Moreover, considering the points in time where participants agreed on unit marks, we found a consistent relationship to computer vision-based touching and untouching events. Specifically, the majority of Ms systematically coincided with a T-U motif, such that Ms followed a touching event and largely co-occurred with a subsequent untouching event. Thus, the most frequently observed motifs were T-U-M (about 27% of the Ms) and T-M-U (about 53% of the Ms). The temporal dispersion of these events in relation to Ms suggested that Ms appeared to be often triggered by a touching event. Thus, the touching events' frequency distribution peaked rather sharply about 260 ms before the M; the untouching events' frequency distribution showed a broader dispersion in time, scattering around the Ms with a mild peak around 90 ms before the M.

It is important to note that T-U sequences were a necessary but not a sufficient condition to bring about an M. That is, if we observed an M, it coincided in most cases (80%) with a T-U motif; but for most (69%) of the T-U motifs, no M was recorded (see Supplementary Table 2). The overall base rate of triplets containing the T-U motif was the highest among all existing triplets, with UTU (41.2%) and TUT (42.4%) being especially frequent. Thus, if participants set a unit mark, they mostly did so in response to a touching event announcing an untouching event, but in many other cases, touching events preceding an untouching event did not trigger a unit marking response. Hence, we can explain the cause for action segmentation in most cases, but also found that humans select one third of these triggering events and disregarded the rest. Note, that Ms could be driven only by T and the relation to U could result from the intervals between T and U. To further investigate this possibility, our explorative findings need to be explicitly tested in future research.

The video content analysis of action phases further elucidated the difference between T-U motifs triggering an M and those that did not. It revealed that, in the first place, Ms announced the object manipulation and the start of the object transport. Less frequently, Ms were placed during the hand transport, during the object transport, and at the end of the object transport. Thus, participants segmented actions particularly during an object manipulation and at the onset of an object transport. These two phases of the observed actions were the only ones that were marked more frequently, almost twice as often, than would have been statistically likely based on the general frequency of occurrence. Notably, object-directed manipulation actions always - and only - consist of two types of phases in variable number and order, i.e., transport and manipulation. Our findings show that at least 80% of human action segmentations can be directly related to the beginning of a transport or the manipulation. Against the backdrop of these novel behavioral findings, we investigated the neural networks associated with the processing of touching and untouching events and their relation to human-determined action segmentation.

Our behavioral findings suggested that touching events are important anchor points of action segmentation, resulting in unit marks distributed around the subsequent untouching event. Touching events themselves, unless they involve grabbing very specific tools in clearly defined contexts, are hardly informative in terms of updating current expectations. Rather, they are mostly points of least predictability of action, as movement comes to a brief halt. Relative to the transport and relative to the phase of manipulation, touchings are therefore more uncertain as the end point of a movement. In our videos, at the time of touching, the now expected manipulation was relatively clearly predictable only in some videos (put cup on saucer), in others not (turn calculator). Such points of lowest predictability were proposed to trigger a visual error signal, initiating upstream areas' updating of the predictive model (Zacks et al., 2011). Fitting this notion, we found increased secondary visual cortex activation comprising cuneus and lingual gyrus pointing to increased exploratory vision and visual gain (Shipp, 2016).

As a counterpart to touching, untouching events terminated the halted movement at touching events and signaled the beginning of the next goal-directed movement. Here, theoretically, competing predictions about potentially upcoming options are retrieved, compared with the actually observed movement at untouching events, and finally disambiguate the observer's expectations. Brain activity at untouching events appeared to reflect these potential processes. On the one hand, activity increased in the anterior dorsal insula (dAI) alerting to a behaviorally important event (Han et al., 2019; Tamber-Rosenau et al., 2018), dorsal anterior cingulate cortex (dACC), which is engaged in saliency detection and attention switching (Han et al., 2019), and finally the inferior frontal junction (IFJ) proposed to subserve transient, dynamic attentional reconfiguration (Sundermann and Pfleiderer, 2012; Xu, 2014). On the other hand, two components that we formerly identified for action segmentation (Schubotz et al., 2012), superior frontal sulcus (SFS) and parahippocampal gyrus (PHG), could now be objectively attributed to the processing of untouching. SFS/PMd serve the selection between alternative competing motor acts based on conditional operations (Petrides, 2005; Tamber-Rosenau et al., 2011). In support of this view, PHG engagement is reliably seen in tasks where contextual associative information is encoded in or retrieved from memory (Aminoff et al., 2013) and is sensitive to stochastic structure of observed events (Amso et al., 2005; Schiffer et al., 2013a; Turk-Browne et al., 2010). Parahippocampal activity extended along the anterior-posterior axis, comprising both posterior and anterior segments which have been related to visuospatial perception and contextual mnemonic processes, respectively (Baumann and Mattingley, 2016). The concurrent engagement of SFS and PHG at untouching events could reflect a comparison between internal model based predicted and actually perceived state changes (Beudel et al., 2016). Summarizing these findings, alertness significantly increases at untouching events, initiating the attentive inspection of the precise hand movement to update expectations and re-focus attention for the upcoming action step.

Object manipulation and object transport unfolding after touching signified a new action segment, and were mostly assigned a unit marker response. Considering brain activity arising at the moment in which participants – in the test/retest sessions following the fMRI experiment – would press the response button to indicate a meaningful action segment, we found strong activation restricted to three areas comprising SPL, IPL, and lateral occipitotemporal cortex. The latter two areas indicate processing of objects, especially in the visuotactile domain, and their manipulation (Creem-Regehr, 2009; Grill-Spector et al., 2001; Lingnau and Downing, 2015), while SPL is involved in vision for action (Gamberini et al., 2020) and, particularly relevant for the present findings, in controlling of all phases of prehension during reach-to-grasp actions (Fattori et al., 2017) as well as observation of reaching/grasping during object manipulation (Wurm et al., 2017). Against the backdrop of the functional profiles of IPL, SPL and LOC, it shows that post-fMRI unit marking coincides with the posterior brain being massively tuned to the analysis of the unfolding step in object manipulation.

Using fMRI and computer vision to investigate human action segmentation was motivated by the suggestion that relying solely on the traditional approach of unit marking behavior does not necessarily tell us which segmental structure the brain processes when we observed actions. Obviously, the brain's ability to recognize and learn statistical structures in stimuli need not be accompanied by our ability to report these structures explicitly (Fiser et al., 2010; Perruchet and Pacton, 2006). The present findings corroborate our assumption, showing that individuals' unit marker responses were tightly bound to T-U motifs, whereas only one third of all T-U motifs triggered a unit marking response. These T-U motifs predominantly indicated object manipulation and the start of object transport. Brain responses for objective and subjective events were clearly distinguishable, and the functional profiles of the activated areas suggested that these events were meaningful and can be interpreted in the context of model updating. Untouching events, and not only those which specifically follow a touching event in a T-U motif, denote action segments as processed by the brain more objectively than human unit marking behavior can do. While to the brain, untouching is informative with regard to the unfolding movement in either case, individuals focused on the moment in which the hand grasped the object to initiate the object manipulation or transport, while occasions for untouching, such as hand-to-object transport, were not considered.

Touching and untouching relations can be reliably detected by computer vision without any need to (train to) identify specific objects (e.g., a pencil) and relate them to typical kinds of manipulation (e.g., writing, drawing). Event segmentation has been shown to be fundamental to how children make sense of the world (Levine et al., 2019) and, speculatively, detecting touching relations could be a very simple way for the baby brain to analyze structure in actions, and learn to recognize recurrent meaningful units way before knowing what we typically do with objects. However, we also know that everyday objects that are familiar to us are strongly associated with certain actions, and this knowledge efficiently modulates the observer's expectation of an action (El-Sourani et al., 2019, 2018; Gupta et al., 2007; Hrkać et al., 2015; Schubotz et al., 2014). Therefore, it would be very important and exciting to investigate what influence this object knowledge has on the segmentation of observed actions.

An important limitation to the generalizability of our results and interpretation concerns the nature of the stimuli used. Our videos were short, discrete, and consisted only of an actress at a table manipulating an object. In contrast, action perception in real life occurs in continuous and more complex contexts. We know from previous studies that the space in which an action is observed (Wurm et al., 2012; Wurm and Schubotz, 2012), the identity of the actor (Hrkać et al., 2013), and contextual objects (El-Sourani et al., 2019, 2018) all have an impact on the brain activity of the action observer. Whether our results are transferable to realistic situations therefore needs to be tested in further studies with more realistic, ecologically valid stimuli.

4.1. Conclusion

Whether we observe actions, listen to music, or hear speech, we easily recognize structure in continuous stimuli. In the present study, using behavioral measures and brain activity, we identified sequential touching relations as a reliable and objective basis for segmenting observed object manipulation. Our findings offer interesting potential applications, for instance, in human-machine interaction, by allowing the machine to make reliable predictions about the way people understand action structures. This information can also help optimizing training protocols used to restore function in stroke patients.

Author contributions

All authors read and approved the final manuscript.

Funding

This work was supported by the German Research Foundation (DFG) [grant numbers SCHU 1439/8-1, WO 388/13-1].

Data availability

Behaviorally determined and objective event data supporting the findings of this study have been deposited in an OSF repository, as well as the source data underlying Fig. 4 and the unit marking test retest raw data (accession code: https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9).

Unthresholded statistical maps of all reported and visualized fMRI contrasts in the manuscript have been deposited on NeuroVault (accession code: https://neurovault.org/collections/8736). The entire stimulus material is available via the Action Video Corpus Muenster (AVI-COM, https://www.uni-muenster.de/IVV5PSY/AvicomSrv/).

The raw fMRI data and the raw SEC time point extraction data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code for the automated extraction of time points of SEC events is available from the corresponding author upon reasonable request. A demo source code of automated extraction that corresponds to the example shown in Fig. 1 can be downloaded from the OSF repository (accession code: https://osf.io/jbwkq/?view_only=e07e36461db248d281597d44c0f83cb9).

Credit authorship contribution statement

Jennifer Pomp: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Nina Heins:** Methodology, Writing – review & editing. **Ima Trempler:** Formal analysis, Writing – review & editing. **Tomas Kulvicius:** Software, Formal analysis, Writing – original draft, Visualization. **Minija Tamosiunaite:** Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing. **Falko Mecklenbrauck:** Formal analysis, Writing – review & editing. **Moritz F. Wurm:** Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Florentin Wörgötter:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Ricarda I. Schubotz:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Acknowledgments

The authors are especially grateful to Mina-Lilly Shibata and Simon Reich for their help with stimulus material recording. Furthermore, we thank Theresa Eckes, Katharina Thiel, and Alina Eisele for their assistance during action video compilation and Monika Mertens for her help during data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118534.

References

Aguirre, G.K., Mattar, M.G., Magis-Weinberg, L., 2011. De Bruijn cycles for neural decoding. Neuroimage 56 (3), 1293–1300. doi:10.1016/j.neuroimage.2011.02.005.

Ahlheim, C., Schiffer, A.-M., Schubotz, R.I., 2016. Prefrontal cortex activation reflects efficient exploitation of higher-order statistical structure. J. Cogn. Neurosci. 28 (12), 1909–1922. doi:10.1162/jocn.

Ahlheim, C., Stadler, W., Schubotz, R.I., 2014. Dissociating dynamic probability and predictability in observed actions: an fMRI study. Front. Hum. Neurosci. 8 (May), 1–13. doi:10.3389/fnhum.2014.00273.

Aksoy, E.E., Abramov, A., Dörr, J., Ning, K., Dellen, B., Wörgötter, F., 2011. Learning the semantics of object-action relations by observation. Int. J. Rob. Res. 30 (10), 1229–1249. doi:10.1177/0278364911410459.

Aminoff, E.M., Kestutis, K., Bar, M., 2013. The role of the parahippocampal cortex in cognition. Trend. Cogn. Sci. 17 (8), 379–390. doi:10.1016/j.tics.2013.06.009.

Amso, D., Davidson, M.C., Johnson, S.P., Glover, G., Casey, B.J., 2005. Contributions of the hippocampus and the striatum to simple association and frequency-based learning. Neuroimage 27 (2), 291–298. doi:10.1016/j.neuroimage.2005.02.035.

Antony, J.W., Harthorne, T.H., Pomeroy, K., Gureckis, T.M., Hasson, U., McDougle, S.D., Norman, K.A., 2020. Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. Neuron 109 (2), 377–390.

Aslin, R.N., 2017. Statistical learning: a powerful mechanism that operates by mere exposure. Wiley Interdiscip. Rev. 8 (1–2), 1–7. doi:10.1002/wcs.1373.

Avrahami, J., Kareev, Y., 1994. The emergence of events. Cognition 53, 239–261. Retrieved from https://www.academia.edu/download/8537960/planners . perspective on art and culture - summer 2010 issue.pdf#page=22.

Baumann, O., Mattingley, J.B., 2016. Functional organization of the parahippocampal cortex: dissociable roles for context representations and the perception of visual scenes. J. Neurosci. 36 (8), 2536–2542. doi:10.1523/JNEUROSCI.3368-15.2016.

Beudel, M., Leenders, K.L., de Jong, B.M., 2016. Hippocampus activation related to 'real-time' processing of visuospatial change. Brain Res. 1652 (May), 204–211. doi:10.1016/j.brainres.2016.10.010.

Botvinick, M., Plaut, D.C., 2004. Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. Psychol. Rev. 111 (2), 395. doi:10.1037/0033-295x.111.2.395, https://doi.org/https://doi.org/.

Brandman, T., Malach, R., Simony, E., 2021. The surprising role of the default mode network in naturalistic perception. Commun. Biol. 4 (1), 1–10. doi:10.1038/s42003-020-01602-z.

Byrne, R.W., Russon, a E, 1998. Learning by imitation: a hierarchical approach. Behav. Brain Sci. 21 (5), 667–684. doi:10.1017/S0140525X98001745, discussion 684-721.

Clewett, D., Davachi, L., 2017. The ebb and flow of experience determines the temporal structure of memory. Curr. Opin. Behav. Sci. 17, 186–193. doi:10.1016/j.cobeha.2017.08.013.

Colder, B., 2011. Emulation as an Integrating Principle for Cognition. Front. Hum. Neurosci. Retrieved from http://www.frontiersin.org/Journal/Abstract.aspx?s=537&name=human_neuroscience&ART_DOI=10.3389/fnhum.2011.00054.

Creem-Regehr, S.H., 2009. Sensory-motor and cognitive functions of the human posterior parietal cortex involved in manual actions. Neurobiol. Learn. Mem. 91 (2), 166–171. doi:10.1016/j.nlm.2008.10.004.

Csibra, G., Gergely, G., 2007. Obsessed with goals": functions and mechanisms of teleological interpretation of actions in humans. Acta Psychol. (Amst) 124 (1), 60–78. doi:10.1016/j.actpsy.2006.09.007.

El-Sourani, N., Trempler, I., Wurm, M.F., Fink, G.R., Schubotz, R.I., 2019. Predictive impact of contextual objects during action observation: evidence from functional magnetic resonance imaging. J. Cogn. Neurosci. 32 (2), 326–337. doi:10.1162/jocn_a_01480.

El-Sourani, N., Wurm, M.F., Trempler, I., Fink, G.R., Schubotz, R.I., 2018. Making sense of objects lying around: how contextual objects shape brain activity during action observation. Neuroimage 167 (June), 429–437. doi:10.1016/j.neuroimage.2017.11.047.

Fattori, P., Breveglieri, R., Bosco, A., Gamberini, M., Galletti, C., 2017. Vision for prehension in the medial parietal cortex. Cereb. Cortex 27 (2), 1149–1163. doi:10.1093/cercor/bhv302.

Fiser, J., Berkes, P., Orbán, G., Lengyel, M., 2010. Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn. Sci. (Regul. Ed.) 14 (3), 119–130. doi:10.1016/j.tics.2010.01.003.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-.P, Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2 (4), 189–210. doi:10.1002/hbm.460020402.

Gamberini, M., Passarelli, L., Fattori, P., Galletti, C., 2020. Structural connectivity and functional properties of the macaque superior parietal lobule. Brain Struct. Funct. 225 (4), 1349–1367. doi:10.1007/s00429-019-01976-9.

Gershman, S.J., Radulescu, A., Norman, K.A., Niv, Y., 2014. Statistical computations underlying the dynamics of memory updating. PLoS Comput. Biol. 10 (11), e1003939. doi:10.1371/journal.pcbi.1003939.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S.S., Maumet, C., … Margulies, D.S., 2015. NeuroVault.Org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* 9 (APR), 1–9. doi:10.3389/fninf.2015.00008.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., Prinz, W., 2007. Predicting point-light actions in real-time. Neuroimage 36 (SUPPL. 2). doi:10.1016/j.neuroimage.2007.03.017.

Grill-Spector, K., Kourtzi, Z., Kanwisher, N., 2001. The lateral occipital complex and its role in object recognition. Vision Res. 41 (10–11), 1409–1422. doi:10.1016/S0042-6989(01)00073-6.

Gupta, A., Davis, L.S., … Park, C., 2007. Object detection action object graphical model objects in action : an approach for combining action understanding and object perception. *Comput. Vis. Pattern Recognit.*. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270329.

Halchenko, Y.O., Hanke, M., 2012. Open is not enough. Let's take the next step: an integrated, community-driven computing platform for neuroscience. Front. Neuroinform. 6, 22. doi:10.3389/fninf.2012.00022.

Han, S.W., Eaton, H.P., Marois, R., 2019. Functional fractionation of the cingulo-opercular network: alerting insula and updating cingulate. Cereb. Cortex 29 (6), 2624–2638. doi:10.1093/cercor/bhy130.

Hard, B.M., Recchia, G., Tversky, B., 2011. The shape of action. J. Exp. Psychol. 140 (4), 586–604. doi:10.1037/a0024310.

Hohwy, J., Hebblewhite, A., Drummond, T., 2021. Events, event prediction, and predictive processing. Top. Cogn. Sci. 13 (1), 252–255. doi:10.1111/tops.12491.

Hrkać, M., Wurm, M.F., Kühn, A.B., Schubotz, R.I., 2015. Objects mediate goal integration in ventrolateral prefrontal cortex during action observation. PLoS ONE 10 (7). doi:10.1371/journal.pone.0134316.

Hrkać, M., Wurm, M.F., Schubotz, R.I., 2013. Action observers implicitly expect actors to act goal-coherently, even if they do not: an fMRI study. Hum. Brain Mapp. 35 (5), 2178–2190. doi:10.1002/hbm.22319.

Kilner, J.M., Friston, K.J., Frith, C.D., 2007. Predictive coding: an account of the mirror neuron system. Cogn. Process 8 (3), 159–166. doi:10.1007/s10339-007-0170-2.Predictive.

Kilner, J.M., Vargas, C., Duval, S., Blakemore, S.-J., Sirigu, A., 2004. Motor activation prior to observation of a predicted movement. Nat. Neurosci. 7 (12), 1299–1301. doi:10.1038/nn1355.

Kosie, J.E., Baldwin, D., 2019. Attentional profiles linked to event segmentation are robust to missing information. Cogn. Res. 4 (8). doi:10.1186/s41235-019-0157-4.

Kurby, C., Zacks, J.M., 2008. Segmentation in the perception and memory of events. Trends Cogn. Sci. (Regul. Ed.) 12 (2), 72–79. doi:10.1016/j.tics.2007.11.004.

Levine, D., Buchsbaum, D., Hirsh-Pasek, K., Golinkoff, R.M., 2019. Finding events in a continuous world: a developmental account. Dev. Psychobiol. 61 (3), 376–389. doi:10.1002/dev.21804.

Lingnau, A., Downing, P.E., 2015. The lateral occipitotemporal cortex in action. Trends Cogn. Sci. (Regul. Ed.) 19 (5), 268–277. doi:10.1016/j.tics.2015.03.006.

Newtson, D., 1973. Attribution and the unit of perception of ongoing behavior. J. Pers. Soc. Psychol. 28 (1), 28–38. doi:10.1037/h0035584, https://doi.org/https://psycnet.apa.org/doi/.

Newtson, D., Engquist, G., 1976. The perceptual organization of ongoing behavior. J. Exp. Soc. Psychol. 12 (5), 436–450. doi:10.1016/0022-1031(76)90076-7.

Newtson, D., Engquist, G.A., Bois, J., 1977. The objective basis of behavior units. J. Pers. Soc. Psychol. 35 (12), 847–862. doi:10.1037/0022-3514.35.12.847.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9 (1), 97–113. doi:10.1016/0028-3932(71)90067-4.

Perruchet, P., Pacton, S., 2006. Implicit learning and statistical learning: one phenomenon, two approaches. Trend. Cogn. Sci. (Regul. Ed.) 10 (5), 233–238. doi:10.1016/j.tics.2006.03.006.

Petrides, M., 2005. Lateral prefrontal cortex: architectonic and functional organization. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 360 (1456), 781–795. doi:10.1098/rstb.2005.1631.

Schiffer, A.-.M., Ahlheim, C., Ulrichs, K., Schubotz, R.I., 2013a. Neural changes when actions change: adaptation of strong and weak expectations. Hum. Brain Mapp. 34 (7), 1713–1727. doi:10.1002/hbm.22023.

Schiffer, A.-.M., Ahlheim, C., Ulrichs, K., Schubotz, R.I., 2013b. Neural changes when actions change: adaptation of strong and weak expectations. Hum. Brain Mapp. 34 (7), 1713–1727.

Schiffer, A.-.M., Krause, K.H., Schubotz, R.I., 2013. Surprisingly correct: unexpectedness of observed actions activates the medial prefrontal cortex. Hum. Brain Mapp. 000. doi:10.1002/hbm.22277.

Schubotz, R.I., Korb, F.M., Schiffer, A.-M.A.-M., Stadler, W., von Cramon, D.Y., 2012. The fraction of an action is more than a movement: neural signatures of event segmentation in fMRI. Neuroimage 61 (4), 1195–1205. doi:10.1016/j.neuroimage.2012.04.008.

Schubotz, R.I., Wurm, M.F., Wittmann, M.K., von Cramon, D.Y., 2014. Objects tell us what action we can expect: dissociating brain areas for retrieval and exploitation of action knowledge during action observation in fMRI. Front. Psychol. 5, 636. doi:10.3389/fpsyg.2014.00636, https://doi.org/https://doi.org/.

Shin, Y.S., DuBrow, S., 2021. Structuring Memory Through Inference-Based Event Segmentation. Top. Cogn. Sci. 13 (1), 106–127. doi:10.1111/tops.12505.

Shipp, S., 2016. Neural Elements for Predictive Coding. Front. Psychol. 7 (November), 1792. doi:10.3389/FPSYG.2016.01792.

Speer, N.K., Swallow, K.M., Zacks, J.M., 2003. Activation of human motion processing areas during event perception. Cogn. Affect. Behav. Neurosci. 3 (4), 335–345. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15040553 .

Sridharan, D., Levitin, D.J., Chafe, C.H., Berger, J., Menon, V., 2007. Neural dynamics of event segmentation in music: converging evidence for dissociable ventral and dorsal networks. Neuron 55 (3), 521–532. doi:10.1016/j.neuron.2007.07.003.

Stadler, W., Schubotz, R.I., von Cramon, D.Y., Springer, A., Graf, M., Prinz, W., 2011. Predicting and memorizing observed action: differential premotor cortex involvement. Hum. Brain Mapp. 32 (5), 677–687. doi:10.1002/hbm.20949, Retrieved from http://dx.doi.org/.

Sundermann, B., Pfleiderer, B., 2012. Functional connectivity profile of the human inferior frontal junction: involvement in a cognitive control network. BMC Neurosci. 13 (1), 1. doi:10.1186/1471-2202-13-119.

Swallow, K.M., Kemp, J.T., Candan Simsek, A., 2018. The role of perspective in event segmentation. Cognition 177 (August), 249–262. doi:10.1016/j.cognition.2018.04.019.

Swallow, K.M., Zacks, J.M., Abrams, R., 2009. Event boundaries in perception affect memory encoding and updating. J. Exp. Psychol. Gen. 138 (2), 236–257. doi:10.1037/a0015631.

Tamber-Rosenau, B.J., Asplund, C.L., Marois, R., 2018. Functional dissociation of the inferior frontal junction from the dorsal attention network in top-down attentional control. J. Neurophysiol. 120 (5), 2498–2512. doi:10.1152/jn.00506.2018.

Tamber-Rosenau, B.J., Esterman, M., Chiu, Y.-.C., Yantis, S., 2011. Cortical mechanisms of cognitive control for shifting attention in vision and working memory. J. Cogn. Neurosci. 23 (10), 2905–2919.

Tobia, M.J., Iacovella, V., Davis, B., Hasson, U., 2012. Neural systems mediating recognition of changes in statistical regularities. Neuroimage 63 (3), 1730–1742. doi:10.1016/j.neuroimage.2012.08.017.

Tootell, R.B.H., Reppas, J.B., Kwong, K.K., Malach, R., Born, R.T., Brady, T.J., … Belliveau, J.W., 1995. Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* 15 (4), 3215–3230. doi:10.1523/jneurosci.15-04-03215.1995.

Turk-Browne, N.B., Scholl, B.J., Johnson, M.K., Chun, M.M., 2010. Implicit perceptual anticipation triggered by statistical learning. J. Neurosci. 30 (33), 11177–11187. doi:10.1523/JNEUROSCI.0858-10.2010.

Wörgötter, F., Aksoy, E.E., Krüger, N., Piater, J., Ude, A., Tamosiunaite, M., 2013. A simple ontology of manipulation actions based on hand-object relations. IEEE Trans. Auton. Ment. Dev. 5 (2), 117–134.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited — Again. Neuroimage doi:10.1006/nimg.1995.1023.

Wu, D.H., Bulut, T., 2020. The contribution of statistical learning to language and literacy acquisition. In: Psychology of Learning and Motivation - Advances in Research and Theory, 72. Academic Press Inc., pp. 283–318. doi:10.1016/bs.plm.2020.02.001.

Wurm, M.F., Caramazza, A., Lingnau, A., 2017. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. J. Neurosci. 37 (3), 562–575. doi:10.1523/JNEUROSCI.1717-16.2017.

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 (7), 1548–1559. doi:10.1162/jocn_a_00232.

Wurm, M.F., Schubotz, R.I., 2012. Squeezing lemons in the bathroom: contextual information modulates action recognition. Neuroimage 59 (2), 1551–1559. doi:10.1016/j.neuroimage.2011.08.038.

Xu, Y., 2014. Inferior frontal junction biases perception through neural synchrony. Trend. Cogn. Sci. 18 (9), 447–448. doi:10.1016/j.tics.2014.06.001.

Zacks, J.M., Kurby, C.a, Eisenberg, M.L., Haroutunian, N., 2011. Prediction error associated with the perceptual segmentation of naturalistic events. J. Cogn. Neurosci. 23 (12), 4057–4066. doi:10.1162/jocn_a_00078.

Zacks, J.M., Swallow, K.M., Vettel, J.M., McAvoy, M.P., 2006. Visual motion and the neural correlates of event perception. Brain Res. 1076 (1), 150–162. doi:10.1016/j.brainres.2005.12.122.