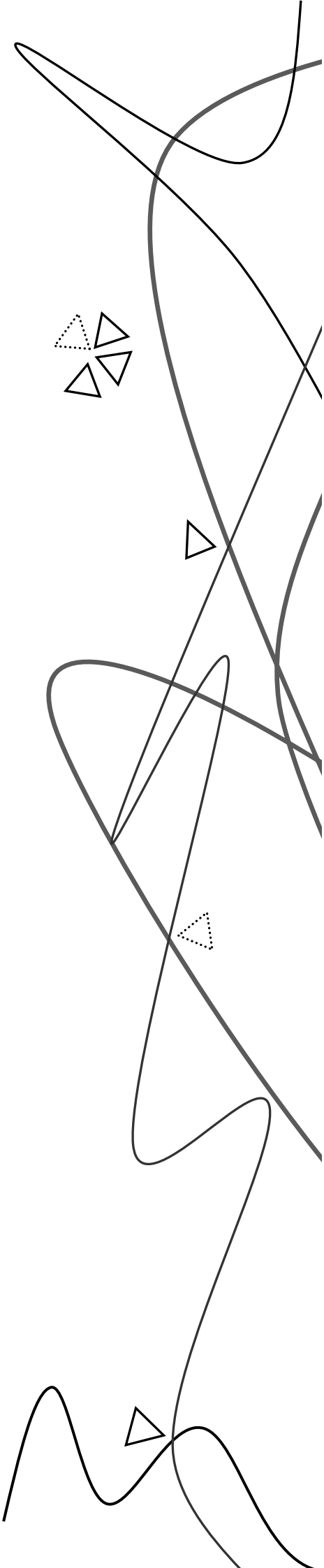

Universität Münster

Jennifer Pomp

The Perception of Structure in Actions: Neural Signatures of Event Boundaries across Behavior and Computer Vision

2025

Psychologie

The Perception of Structure in Actions: Neural Signatures of Event Boundaries across Behavior and Computer Vision

Inaugural-Dissertation
zur Erlangung des Doktorgrades
im Fachbereich Psychologie und
Sportwissenschaft der Universität Münster

Vorgelegt von
Jennifer Pomp
aus Münster
- 2025 -

ORCID 0000-0002-8577-7948

Dekanin:	Prof. Dr. Ricarda I. Schubotz
Erste Gutachterin:	Prof. Dr. Ricarda I. Schubotz
Zweite Gutachterin:	Prof. Dr. Pienie Zwitserlood
Tag der mündlichen Prüfung:	10. Juli 2025
Tag der Promotion:	10. Juli 2025

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported and contributed to my journey throughout this PhD. Without the encouragement, guidance, and companionship of so many, this work would not have been possible.

First and foremost, I would like to thank my supervisor, Professor Ricarda I. Schubotz, for giving me the opportunity to pursue a PhD in cognitive neuroscience. Her guidance, constructive feedback, and encouragement have been invaluable throughout this process. She provided me with the freedom to explore new ideas and make this work my own. I appreciate her flexibility, granting me the autonomy to balance the demands of my academic work with my personal life.

Furthermore, I would like to express my gratitude to my co-supervisor and collaborators, Professor Florentin Wörgötter and Professor Minija Tamosiunaite, for their valuable contributions to this research. Their expertise, commitment and perspective in the field of robotics have made this project possible. In addition, I thank my co-supervisor Professor Pienie Zwitserlood for her interest in my work and for giving advice whenever needed.

To my amazing co-authors, thank you for your collaborative efforts and valuable contributions to the research presented in this thesis. A special thanks goes to Professor Moritz Wurm who always had a sympathetic ear for methodological challenges and reliably gave sound advice from afar.

I also want to thank my colleagues from the biological psychology lab for their camaraderie, their collaboration, intellectual exchange, and constant support. The insightful (and often very humorous) discussions and their collective wisdom have been incredible sources of support and inspiration. Special thanks go to Dr. Ima Trempler, Dr. Nina Heins, Dr. Nadiya El-Sourani, Dr. Lena Leeners, Dr. Benjamin Jainta, Dr. Anoushiravan Zahedi, and Rosari Naveena Selvan, who have made this experience all the more rewarding. I am grateful for the

opportunity to have worked alongside such talented individuals. Not forgetting Monika Mertens and Jutta Linke for the nice company during long hours of MRI measurements and the valuable assistance with the administrative aspects of my work. In this context, I would also like to thank all the student assistants who helped me during this journey, for their support and dedication.

In addition, I wish to express my sincere gratitude to my friends, thank you for your belief in me and for being a source of ease, laughter, and balance during the ups and downs of this journey. Your understanding, encouragement, and moments of distraction have kept me going.

Furthermore, I owe an immeasurable amount of gratitude to my family who always accepted my decisions and supported me in realizing them. To my mom: *Danke für deine bedingungslose Liebe und Unterstützung auf meinem Weg. Ohne dich wäre ich nicht da, wo ich heute bin*. To my brothers and sisters-in-law, your love, patience, friendly side blows and sarcasm mean a lot to me. To my husband, words cannot fully express how much I appreciate your patience, love and understanding throughout this journey. You always believed in me, and I am endlessly grateful that you took on responsibilities when I needed to focus. Finally, I am deeply thankful to my daughters. Your joy, curiosity, and boundless energy have enriched my life in an indescribable way. You taught me balance, resilience, and enjoying the little things in life. Seeing you grow has been a constant source of delight, reminding me of what really matters.

This thesis is as much a reflection of your support and belief in me as it is of my own persistence and hard work. I am truly blessed to have you all in my life.

Contents

Summary

In everyday perception we automatically draw structure out of the continuous stream of information that we encounter. Chunked into events, we can process these units of experience offering the structure for memory and prediction. Event perception is a fundamental cognitive process shaping our experience. The mechanisms underlying the detection of boundaries between events continue to be a central focus in cognitive (neuro)science. While previous studies mainly concentrated on participants' behavioral event annotations, little is known about the objective stimulus features driving the segmentation behavior. To address this, the current thesis employed computer vision methods extracting stimulus characteristics to derive objective event boundaries as well as action categories. The aim was to examine objective event boundaries, their value for understanding subjective event boundaries and the neural underpinnings of both. Object-related action associations were considered as a modulating factor. In addition, the neural representation of objective and subjective action categories was investigated.

For this purpose, two experiments, each consisting of three sessions, were conducted. In the first session, participants passively observed short videos of object-directed actions during the MRI scan, to segment these manually in the second and third experimental session (test-retest). Subsequently, the participants performed a multi-arrangement task in which they spatially organized actions according to their similarity. In the first experiment, the actions were directed at commonplace items (e.g., a calculator, a cup, or a piggy bank) while in the second experiment, formed pieces of dough were manipulated. The actions remained the same over experiments, as well as the experimental tasks and procedures, but the manipulated items varied in the strength of object-action associations. For the analyses, subjective event boundaries were determined on group level based on consistent individual segmentations and objective event boundaries were extracted by computer vision algorithms. To derive the latter,

relational changes between objects in the form of touchings and untouchings between hands, objects, and the ground were determined and coded as a sequence to describe the corresponding action category. These (un)touching sequences have proven highly useful for robots to recognize human actions and execute these actions itself. Therefore, they were considered relevant objective event boundary candidates, to which human event structure processing could also relate.

Study I used the fMRI and behavioral segmentation data of the first experiment to investigate whether (un)touchings are a meaningful supplement or reference point to subjective event boundaries and how they contribute to understanding neural event structure processing in object-directed action observation. Both subjective and objective event boundaries showed definable underlying neural activation patterns, and the temporal co-occurrence of the boundaries suggested a key role of objective boundaries for identifying events' limits.

Based on the data of both experiments, Study II investigated the modulating effect of object-action associations on the segmentation behavior and the neural processing at subjective and objective event boundaries. The results confirmed objective boundaries to be meaningful anchor points for subjective boundaries with behavioral annotations being even closer to objective boundaries when object-associated knowledge was limited. Furthermore, they revealed a significant effect of association strength on underlying brain processing. At untouchings, limited object-action associations were accompanied by increased biological motion processing and strong associations, in contrast, by increased contextual information processing. At the same time, activity in the anterior inferior parietal lobule (aIPL) increased for weak object-action associations which was interpreted as mirroring an unrestricted number of candidate actions for predicting the unfolding action.

Finally, Study III used a representational similarity analysis of the fMRI data from both experiments to investigate whether objective action categories are represented in the neural

processing patterns and whether they are related to subjective action spaces derived from the multi-arrangement task. Subjective action categories were associated with a broad bilateral network while objective action categories were selectively associated with the representational profile of the aIPL. A significant relationship between the two action spaces emerged only when object information was limited.

Collectively, these findings indicate that objective event boundaries are a meaningful addition to subjective boundaries in understanding the processing of object-directed actions. They provide objective anchor points for segmenting actions behaviorally and aid in disentangling the neural signatures of event structure. Concerning the neural profiles across experiments, subjective event boundaries were mainly motion-driven and low-level visual inspection of the scene intensified at points of touching. Remarkably, the points of untouching were revealed to be important for attentional recalibration, memory encoding and predicting the upcoming action step. Furthermore, it was at this exact point that the strength of object-action associations became evident. Thus, the points of untouching appear to play a significant role. The current work further elaborated on the role of the aIPL in action observation. The aIPL is suggested to serve a critical function in predicting object-related actions based on object-associated action knowledge and in representing objective action categories.

This thesis offers a new perspective on event structure perception through objective, stimulus-derived event boundaries and action categories. The results suggest important implications for neurorehabilitation settings as they could help optimize training protocols. Similarly, the results could inform the development of robotic systems that support patients with motor impairments and enhance human-robot cooperation. This thesis lays the groundwork for more detailed investigations into neural event structure processing, with the aim of ultimately understanding this core capacity that fundamentally shapes our experience.

List of Original Publications

This thesis is based on the following original research articles:

Study I — Pomp, J., Heins, N., Trempler, I., Kulvicius, T., Tamosiunaite, M., Mecklenbrauck, F., Wurm, M.F., Wörgötter, F., & Schubotz, R. I. (2021). Touching events predict human action segmentation in brain and behavior. *Neuroimage*, *243*, 118534. https://doi.org/10.1016/j.neuroimage.2021.118534

Study II — Pomp, J., Garlichs, A., Kulvicius, T., Tamosiunaite, M., Wurm, M. F., Zahedi, A., Wörgötter, F., & Schubotz, R. I. (2024). Action Segmentation in the Brain: The Role of Object–Action Associations. *Journal of cognitive neuroscience*, *36*(9), 1784-1806. https://doi.org/10.1162/jocn_a_02210

Study III — Pomp, J., Wurm, M. F., Selvan, R. N., Wörgötter, F., & Schubotz, R. I. (2025). Touching-untouching patterns organize action representation in the inferior parietal cortex. *Neuroimage*, 310, 121113. https://doi.org/10.1016/j.neuroimage.2025.121113

1 Theoretical and Empirical Background

As we go about our daily lives, we constantly perceive the world through our senses, with a continuous flow of information coming our way. Still, when we are asked to report what we experienced on any given day, we usually report integrated and coherent but discrete episodes. Thus, the continuous information has been chunked into meaningful units. These units of experience (Yates et al., 2023), that have been termed "events", caught a lot of scientific attention during the last decades. The resulting field of event perception research has grown significantly (e.g., Bailey & Smith, 2024; Dubrow, 2024; Radvansky & Zacks, 2011, 2017; Zacks, 2020). It investigates event perception over the lifespan (e.g., Zacks et al., 2006; Zheng et al., 2020) and of various perceptual input formats (e.g., sequences of images: DuBrow & Davachi, 2016; story listening: Kumar et al., 2023; auditory event sounds: Ogg & Slevc, 2019; story reading: Pettijohn & Radvansky, 2016; action videos: Pomp et al., 2021). The study of event representation has typically addressed either the events' properties or the characteristics of their "boundaries" (i.e., the point when one event ends and the subsequent event begins; Yates et al., 2023). Alongside various approaches, cognitive (neuro)scientists consult action observation to understand event boundaries in detail.

When we observe someone performing an action, we perceive the action as a continuous flow of movements. For instance, when someone prepares their breakfast, we see an action unfolding and we can, when asked for, divide the continuous action into distinct meaningful units (i.e., events) and identify these segments or action phases (e.g., *preparing a sandwich*, *making coffee*, etc.). Naturally, we can even do this at different levels of coarseness. For instance, we can recognize the steps of *preparing a sandwich* as: *taking a loaf of bread*, *buttering it*, and *placing a slice of cheese on it*; or zoom in further describing the action steps on a more detailed level (e.g., dividing only *taking a loaf of bread* into distinct segments). This subjective segmentation of observed actions is one focus of action observation research and

intends to understand the perception, processing and storage of sequences of events. One approach to obtain subjective action segments is to ask action observers to indicate action steps by button press during observation. This procedure revealed an intra-individually highly consistent segmentation behavior (Newtson, 1973; Newtson & Engquist, 1976; Zacks, Tversky, et al., 2001). In fact, event segmentation has been shown to be an automatic part of ongoing perceptual processing (Yates et al., 2022; Zacks & Swallow, 2007) and splits continuous input into distinct units or events that are separated by boundaries.

The traditional approach of asking action observers to indicate action steps requires individuals to consciously decide on boundaries (i.e., when to press a button). Thus, this kind of event boundary detection invariably includes an explicit filter. We know from other research areas that we do not necessarily have access to the information that matters for the observer's brain to make sense of the world around it. For example, the phenomenon of implicit learning shows that we extract regularities of the world without a clear awareness of what we know (Perruchet & Pacton, 2006; Williams, 2020; for a review see Stadler & Frensch, 1998). Several studies have shown that people are able to implicitly learn the statistical structure that underlies incoming stimulus streams of observed actions (e.g., Ahlheim et al., 2014; Swallow & Zacks, 2008). When explicitly asked about it, participants are unable to report the learned structure. Hence, subjective report does not necessarily reveal underlying perceptual processes. Derived from this, it can be assumed that there could be regularities in an observed action that mark meaningful event boundaries to the observer's brain but have no relevance on the conscious level. Therefore, the current work distinguishes between observer-labeled event boundaries and stimulus-derived event boundaries and considers both as worth investigating to understand the perception of event structure and underlying neural processes.

Stimulus-based event boundaries can take various forms. Remarkably, regarding the segmentation of narrative events, a computational approach using a large language model

(GPT-3) was recently able to roughly reproduce human event annotations (Michelmann et al., 2025). The model proved to be closer to average behavioral annotations than individual human annotators were. The authors suggest GPT-3 to provide a reasonable solution for automated event annotations. In an intriguing way, this specific case of event boundary detection is neither what I consider observer-labeled nor what I consider stimulus-derived. The former is obvious, whereas the latter requires further consideration. In fact, the model-detected event boundaries are not inherent to the stimulus alone as the large language model is needed to evaluate the narrative in the context of written human language.

In this work, the term “stimulus-derived event boundaries” is used in a specific sense, referring to event boundaries that are independent of an observer’s decision or perception and can be extracted from the stimulus per se. They are objective in nature. They allow, inter alia, a high level of between-subject comparability for neuroscientific investigations as the perceptual input at boundaries is constant across participants[1]. In addition, from a multidisciplinary perspective, this kind of event boundary is of growing interest to computer vision and artificial intelligence (AI) for visual event detection. The task of identifying events in visual data is used in automated systems to analyze action sequences, such as in video surveillance systems (for a review see Jebur et al., 2023), autonomous driving (for a review see Xiao et al., 2023), video analytics (e.g., Canel et al., 2019) and sports broadcasting (e.g., Xu et al., 2006). The term “observer-labeled event boundaries”, in contrast, refers to the subjective unit annotations made by the participants during the observation of an action.

In cognitive neuroscience, event perception is not merely a perceptual function – it is the core mechanism through which the brain constructs reality, providing the temporal scaffolding for memory and prediction. This thesis seeks to advance our understanding of event perception and to establish objective anchor points for the segmentation of events. To this end, the present

[1] If individual differences in attention processes are disregarded.

work focuses on subjective (i.e., observer-labeled) and objective (i.e., stimulus-derived) event boundaries in action observation and their neural processing. Furthermore, the scope of this work extends to the representation of the action categories that can be determined both subjectively (i.e., based on participants' ratings) and objectively (i.e., based on computer vision). To complement these, object-action associations are examined as a modulating factor. This factor is explained in more detail below. The following sections will provide an overview of visual action perception, including events and predictive processing, and examine the underlying neural processes. In addition, computational action representation in neuroscience will be addressed.

1.1 Actions and Action Perception

Action observation research encompasses a broad spectrum of actions that are investigated. Within the field, these actions can largely be divided into whole body movements (e.g., swimming, walking) and hand movements, while some studies also show movements of the feet or face (cf. Caspers et al., 2010). Dima et al. (2024) recently demonstrated that partcipants' similarity judgments reflected a shared organization of actions across videos and sentences. This organization was mainly determined by the target of the action (i.e., whether the action was directed towards an object, another person, or the self) which validates the distinction between actions suggested by the field. In addition to the type of action, the context in which an action is presented varies considerably across studies. It ranges from tightly controlled, purpose-designed stimuli to more naturalistic cinematic content, and extending further to immersive virtual environments (Pooja et al., 2024).

1.1.1 Object-directed Actions

Hand actions can be divided into transitive actions, which are directed towards objects in the peripersonal space (e.g., grasping a cup), and intransitive actions, which do not involve an object (e.g., waving a hand). Transitive actions are also termed "object-directed" (in contrast to "object-unrelated"). They are the focus of this thesis. The most important (i.e., primary) sources of information in object-directed actions are the movements and the objects involved (Wurm & Schubotz, 2017). Concerning the latter, the familiarity of an object is a crucial factor. Everyday objects that are familiar to us are strongly associated with actions that we typically perform with them, and this knowledge modulates our expectations regarding the upcoming action (El-Sourani et al., 2018, 2019; Hrkać et al., 2015; Kalénine et al., 2016; Schiffer et al., 2012; Schubotz et al., 2014; Schubotz, 2015). In the same way, potential interactions between objects (if multiple objects are present) shape action perception as they have been shown to be extracted automatically and directly (S. Xu & Heinke, 2017). The effect of the implied actions between objects can selectively be reduced using online repetitive transcranial magnetic stimulation (rTMS; S. Xu et al., 2017). Consequently, it is likely that object-associated action knowledge modulates event structure perception. Its modulating effect on action segmentation and prediction is one of the main aspects investigated in this thesis.

1.1.2 Events and Their Boundaries

Observing behaviors that unfold over time and segmenting these actions into separate events raises the question of how those events can be defined. An event may be thought of as a distinct unit of individual experience, which organizes continuous perceptual input into mental units that we can label, remember and search for in memory. Some authors suggest that events are temporal building blocks used by our cognitive system, just as objects are spatial building blocks (see e.g., Tversky et al., 2004). There is, however, no consensus in cognitive (neuro)science or psychology on what events are and what they are not (Reilly et al., 2025; Yates

et al., 2023). Nevertheless, event perception is the focus of various research fields and within this research area, event segmentation drew attention to the boundaries between events (for a review see Zacks, 2020) which are central to this work. Event segmentation paradigms serve the purpose of finding these event boundaries in the continuous stream of sensory input.

More than 50 years ago, using the unit-marking procedure (i.e., an event segmentation paradigm) in several behavioral studies, Newtson and colleagues demonstrated that action observers exhibit an inter-individually variable, but intra-individually highly consistent segmentation behavior when asked to indicate action steps (Newtson, 1973; Newtson & Engquist, 1976). The unit-marking procedure comprises that participants watch an action video and press a button whenever they think one unit ends and another one begins. It is still today a valuable tool to study how individuals perceive and segment actions into discrete units while variants emerged that use not only videos and movies but also slideshows and reading of or listening to stories (Sargent et al., 2015; for a review see Zacks, 2020). Subsequent research revealed that marked action segments resist interruptions (Newtson & Engquist, 1976), missing content (Kosie & Baldwin, 2019) and perspective shifts (Swallow et al., 2018). Furthermore, action stream breakpoints or event boundaries receive increased attention (Hard et al., 2011), are better recognized than other intervals (Pradhan & Kumar, 2022; Swallow et al., 2009), and at boundaries observers are less likely to mind-wander (Faber et al., 2018). Action representations are structured by event boundaries that drive also memory (Ezzyat & Davachi, 2011, 2021; Kurby & Zacks, 2008; Pettijohn et al., 2016; Pradhan & Kumar, 2022; Swallow et al., 2009; Zacks, Speer, et al., 2006) and planning (Zacks et al., 2011) while an attentional focus on individual situational dimensions (as e.g., spatial information) can influence boundary marking (H. R. Bailey et al., 2017; De Soares et al., 2024). A recent work by Sasmita and Swallow (2023) investigated the stability of event boundary agreement within and across groups and demonstrated the reliability of segmentation performance in different experimental setups and sample sizes.

In addition to subjective, observer-labeled event boundaries, some approaches determined objective, data-driven features that relate to event boundaries. For instance, event structures were extracted from movement parameters (Zacks et al., 2009), and participant-judged boundaries were associated with bursts of change in movement features (Hard et al., 2006). Furthermore, stimulus characteristics as the statistical structure can play a role as an objective reference in human action segmentation (Baldwin et al., 2008). These approaches have in common that they aim to ground subjective boundary annotations in objective stimulus features. Magliano and Zacks (2011), in contrast, used another method to study how people perceive the structure of events and investigated the impact of continuity editing[2] in narrative film's segmentation. Watching narrative films is a special variant of action observation as perspectives, locations and agents constantly change and thus the flow of information across shots often shows little similarity to the perceptual input when we observe the real world (Cutting, 2005). Nevertheless, it offers important insights as the study by Magliano and Zacks (2011), for instance, clearly indicated that discontinuity of action was the strongest predictor for a behavioral event boundary, compared to spatial-temporal changes. Brain activation patterns at action discontinuities particularly showed decreased activity at posterior temporal, inferior and superior parietal and dorsal premotor cortex along with increased activation in lateral occipital regions. The reductions in activity were interpreted as attention-driven down-regulation of processing to wait until the parameters of the new scene were established. Hence, Magliano and Zacks (2011) investigated points of discontinuity as objective event boundary candidates. However, this approach is not applicable to the segmentation of uncut video material that is supposed to mimic real world action observation.

[2] Continuity editing is used by filmmakers to evoke a sense of situational (dis)continuity at editing boundaries.

1.1.3 Predictive Action Processing

The framework of predictive coding suggests that the brain constantly generates predictions about sensory input and updates those predictions based on the incoming sensory information. While analogous concepts have been introduced earlier, the currently known concept of predictive coding has been mainly shaped by a few influential works (Clark, 2013; Friston, 2005, 2010; Hohwy, 2013; Rao & Ballard, 1999). Research from various fields continues to expand the reach of this theoretical framework. In contrast to most preceding theories of brain function, the idea is that the brain does not passively receive information but actively predicts the sensory environment. When the incoming information does not match the generated prediction, a prediction error arises, and the internal model gets updated to improve future predictions. This process unfolds across multiple hierarchical levels, so that predictions are passed on top-down and prediction errors are passed on bottom-up. Transmitting only the unpredicted portion of a sensation is metabolically efficient. Importantly, grasping the brain as a prediction machine is not only about perception but also about action as active inference (Clark, 2013; Friston, 2010; Hohwy, 2013). This means that actions can be initiated to actively generate the predicted sensations. Furthermore, Clark (2024) recently elaborated about hacking our own predictive brain to better serve our needs, which shows that the framework of predictive processing is consistently widening.

Regarding action prediction, the predictive coding framework suggests that the brain anticipates the consequences of our own actions. Thus, the brain predicts the sensory input that comes in when we act on the world. This includes proprioception as well as tactile perception when we touch something, visual perception when we see our action, auditory perception when our action produces a sound, olfactory perception when we expect to smell something (e.g., because we grasp a fragrant flower and move it closer to our nose), or gustatory perception. This is essential to plan and execute movements and informs online action

coordination. Kilner et al. (2007) combined the predictive coding account with action observation research to explain the inference of intentions when making sense of others' actions. He proposed that the ability to understand observed action at the abstract level of intentions is encoded in middle temporal and inferior frontal brain regions that predict the most probable intentions and goals of the observed action (Kilner, 2011).

Relating the predictive coding framework to event boundaries, Reynolds et al. (2007) ran simulations to demonstrate that a system can accurately identify event boundaries based on prediction error increases. Zacks (2020) elaborated the fact that there is a close temporal relationship between event segmentations and moments of low predictability. Observers were more likely to identify event boundaries as the course of an action became more unpredictable, and those boundaries were related to enhanced memory and a reduced ability to predict the upcoming event (Huff et al., 2014). Event boundaries corresponded to the point in time where it was more difficult for participants to predict the near future (Zacks et al., 2011) so that participants made better predictions within an event than across event boundaries. Furthermore, predictive eye movements are less prevalent around event boundaries (Eisenberg et al., 2018) suggesting that prediction is vague at this point. These findings suggest that predictability is high during an event and low at event boundaries. Said differently, prediction error is low during events and increases at event boundaries. The increased prediction error at event boundaries then triggers internal model updates (Zacks, 2020). The fundamental assumptions of this approach were formulated in the event segmentation theory (Zacks et al., 2007). It assumes that people construct and maintain representations of the currently unfolding action and predict what will happen next on basis of sensory cues and knowledge structures. Transient errors in the predictions result in the perception of an event boundary. Applied to the example from the beginning of someone preparing their breakfast: after placing a slice of cheese on a loaf of bread to prepare the sandwich, it becomes less predictable what will happen next as several upcoming action steps are possible; the person may take a bite or cut the sandwich in

halves (or do something completely different). The incoming sensory information is used to identify which of the typically expected actions will occur, and the internal model is updated accordingly. There is abundant evidence of empirical findings that are compatible with predictive approaches of action observation (e.g., Cerliani et al., 2022; Kemmerer, 2021; Keysers et al., 2024; Urgen & Saygin, 2020; Zentgraf et al., 2011) and event boundary perception (e.g., Ezzyat & Davachi, 2021; Reagh et al., 2020; Schubotz et al., 2012).

At the same time, there are empirical findings regarding event boundaries that are difficult to reconcile with the predictive coding framework (Yates et al., 2023) and likewise for action observation (Kemmerer, 2021). Alternative theories of event segmentation rely on inferences about what generates an experience so that an event boundary occurs when the inference changes (e.g., Shin & DuBrow, 2021). This allows boundaries to occur independently of perceptual change or low predictability. In addition, Franklin et al. (2020) designed a probabilistic reasoning model to demonstrate that it can produce human-like segmentations of naturalistic data and that its principles are sufficient to explain a wide range of empirical findings. This is an example of computer science modeling human abilities to illustrate which principles may be sufficient, bringing us closer to understanding how event structures could be built and how boundaries could be perceived. However, it remains unresolved which theory provides a more accurate representation of the mechanisms in the brain.

1.2 Neural Action Processing

Functional imaging studies revealed remarkable findings about the neural processing of observed actions. This section describes the neural action observation network as well as the neural signatures of event boundaries and action categories.

1.2.1 The Action Observation Network

Over the last two decades, many neuroimaging studies have assessed the human brain networks underlying action observation (for meta-analyses see Caspers et al., 2010; Hardwick et al., 2018) and its development over the lifespan (Biagi et al., 2016; Lesourd et al., 2023; Morales et al., 2019; Sacheli et al., 2023). The increasing interest in neuroscientific research on action observation can largely be attributed to the discovery of mirror neurons in nonhuman primates. This unique class of neurons responds both during the execution of an action and the observation of someone else performing this action. Mirror neurons were first discovered in macaque area F5 (Di Pellegrino et al., 1992; Rizzolatti et al., 1996) which is the putative homologue of the human premotor cortex. Thereafter, mirror neurons were also found in macaque rostral inferior parietal lobule, which is the putative correspondence to the human anterior inferior parietal lobule (Fogassi et al., 2005).

For ethical and practical reasons, single-cell recordings are not conducted in healthy humans, so noninvasive brain imaging techniques were used to study human action observation. The emerging human action observation network has been found to expand the above-mentioned regions and includes the premotor, parietal and temporo-occipital cortex (Caspers et al., 2010; Gazzola & Keysers, 2009; Hardwick et al., 2018; Kilner, 2011; Lesourd et al., 2023). It has been investigated using various tasks and paradigms which differ, for instance, in terms of the effectors (e.g., hand, foot, or face), instructions (e.g., passive observation or observation to imitate) and the involvement of an object (transitive versus intransitive actions) (Caspers et al., 2010; Hardwick et al., 2018). Given the thesis's focus, the next sections review neuroscientific paradigms of action observation that specifically address action segmentation and categorization.

1.2.2 Neural Signatures of Event Boundaries

Neuroimaging studies employed functional magnetic resonance imaging (fMRI) to investigate the neural signature of behaviorally determined event boundaries compared to non-boundary points. Subcortically, increased activity was observed in the hippocampus (Ben-Yakov & Henson, 2018; Reagh et al., 2020) as well as in the adjacent parahippocampal cortex (Ben-Yakov & Henson, 2018; Reagh et al., 2020; Schubotz et al., 2012). On the cortical level, various regions were also activated at event boundaries, such as the angular gyrus, the visual cortex, the precuneus, the temporoparietal and the occipitotemporal junction, the superior temporal sulcus, posterior medial regions and the superior frontal sulcus (Ben-Yakov & Henson, 2018; Betti et al., 2013; Ezzyat & Davachi, 2011; Reagh et al., 2020; Schubotz et al., 2012; Speer et al., 2003; Zacks et al., 2011; Zacks, Braver, et al., 2001). Evidence regarding the role of these regions in event detection is mixed, with varying degrees of clarity.

Boundary-evoked hippocampal activation has been associated with memory performance (Reagh et al., 2020) and it has been suggested that the increased hippocampal activity could reflect the registration of the preceding event to long-term memory (Ben-Yakov & Henson, 2018). This is in line with the idea that an event boundary segregates the immediate present (that is active in working memory) from preceding events (that are registered in long-term memory) so that the experience of event structure shapes the memory content (Ezzyat & Davachi, 2011; Zacks, 2020). For the cortical activation pattern, the indications are less clear. Angular gyrus activation and superior frontal sulcus activation have been functionally interpreted as engaging spatial attention, and parahippocampal activation as being associated with long-term memory retrieval (Schubotz et al., 2012). At the same time, cortical activation patterns were frequently interpreted referring to their functional or structural affiliation to the hippocampus (Reagh et al., 2020; Schubotz et al., 2012). Increased activation in the lateral occipitotemporal region were mostly related to motion sensitive areas processing movement

features at event boundaries (Speer et al., 2003; Zacks, Braver, et al., 2001; Zacks, Swallow, et al., 2006).

Furthermore, functional connectivity analyses revealed that the interactions between the hippocampus and the posterior medial network (i.e., a default mode subnetwork consisting of the posterior cingulate cortex, the retrosplenial cortex, and the angular gyrus) at event boundary encoding was associated to subsequent successful event recall and the amount of recalled detail after a delay (Barnett et al., 2024). A recent study showed that multivoxel patterns of past events are reactivated at event boundaries modulated by the similarity of their semantic content, which was demonstrated in the hippocampus, medial temporal lobes and posterior medial cortex (Hahamy et al., 2023). Thinking one step ahead, a novel method modeled neuroimaging data directly with a dynamic event segmentation model. In this data-driven approach, Baldassano et al. (2017) discovered brain activation patterns that were associated with the event structure in narrative stimuli and a nested hierarchy from short to long events. Here again, the angular gyrus, the posterior medial cortex, the parahippocampal cortex and the hippocampus played a crucial role. In a similar manner, Yates et al. (2022) applied a computational model to brain activation patterns of infants to identify event signatures. They revealed that infants, in contrast to adults, segment fewer, longer events across the cortical hierarchy. In sum, an emerging body of imaging evidence starts to shed light on the neural processing of event boundaries.

1.2.3 Neural Activation Patterns Associated with Action Categories

Understanding how the brain organizes and differentiates between various types of knowledge is a key challenge in cognitive neuroscience. Prior studies have shown that different semantic categories elicit distinct patterns of neural activation across cortical regions (see e.g., Binder et al., 2009; Malone et al., 2016). To investigate neural activation patterns underlying action spaces, the way of categorization is central. Most studies from the last decade use pre-

determined stimulus taxonomies and behaviorally determined similarity spaces to capture different representational geometries that are compared to neural representational patterns from multivoxel pattern analyses. In general, clear methodological parallels are recognizable from the domain of object recognition (Lingnau & Downing, 2024). Prior to the widespread adoption of multivoxel pattern analyses, imaging studies used univariate contrasts to explore the representation of semantic spaces. For instance, the meta-analysis by Binder et al. (2009), that reviewed the representation of semantic word processing, yielded distinct semantic subsystems and localized action knowledge in left supramarginal gyrus (SMG) and posterior middle temporal gyrus (pMTG). The stimulus format is critical here as Wurm and Caramazza (2019) showed action representation to differ across vision and language. They revealed that frontoparietal areas discriminated observed action scenes and corresponding written sentences, but the decoded representations were overlapping and not generalized across stimulus types. The left lateral posterior temporal cortex, in contrast, encoded generalized action representations. Furthermore, Wurm et al. (2017) identified neural representations of actions to be organized along sociality (i.e., nonsocial vs. social) and transitivity (i.e., object-unrelated vs. object-related) in bilateral lateral occipitotemporal cortex (LOTC). Additionally, they suggested a posterior-anterior gradient in LOTC from concrete to abstract action features. Regarding different abstraction levels, another study by this workgroup demonstrated the inferior parietal and occipitotemporal cortex to code actions at abstract levels and the premotor cortex to code actions at the concrete level only (Wurm & Lingnau, 2015).

The functional role of the parietal cortex in action observation has further been differentiated from the occipito-temporal and premotor cortex. The discriminability between action classes was higher in the parietal cortex, suggesting that action identity is coded in this region (Urgen & Orban, 2021). To summarize, the regions that are generally counted as part of the action observation network are involved in the representation of action spaces to varying degrees and various dimensions have been described to organize actions and their neural

representation. Currently, many outstanding questions are actively addressed in the field to broaden our understanding of neural action representation.

1.3 Computational Models of Action Representation in Neuroscience

Action recognition is important for computer vision since applications like visual surveillance, autonomous driving, human-robot interaction, augmented entertainment and video retrieval are of constantly growing interest and the availability of big data opens up new opportunities. In computer science, action recognition has been extensively investigated in the last decades such that great successes have been achieved in after-the-fact action recognition (i.e., recognition after observing the entire action execution) and, recently, even action prediction (i.e., recognition before action execution is completed) is being pursued (for a review see Kong & Fu, 2022). There are two main challenges in vision-based action recognition, that is, action representation and action classification (Kong & Fu, 2022). Remarkably, there is significant potential for cooperation between computer vision and cognitive neuroscience, making their integration highly valuable for advancing our understanding of visual processing. Modeling brain activation data with objectively and automatically determined stimulus characteristics is just one example illustrating the unique avenue that computer vision provides. The following sections describe how this thesis employed computer vision methods to explore a neuroscientific research question.

1.3.1 Actions Represented as Touching Relations Between Objects

One advantage of computer vision is that it can extract the static and dynamic characteristics of stimuli objectively and automatically. To represent an action based on these, very different approaches exist, and some approaches concentrate on objects and their relationship. For instance, Ji et al. (2020) represent actions in spatio-temporal scene graphs that code the objects and their relative spatial position (e.g., person – sitting on – sofa). This

representation is substantially reduced, and further simplification is possible. The method implemented in the present work represents object manipulations by coding the spatial contact between surfaces (e.g., object one – touching – object two). It constructs a dynamic graph sequence from continuously tracked RGB-D sensor data of action videos (Aksoy et al., 2011; Worgotter et al., 2013). In these graphs, objects build the nodes, and a touching relation is represented by an edge. Topological transitions of such a graph occur whenever objects touch or un-touch and are stored in a transition matrix called the semantic event chain (SEC). Remarkably, this account is model-free and strictly stimulus-driven: It does not distinguish between hands, objects, or the ground, nor does it require any functional or semantic knowledge about objects as it does not identify them.

Therefore, in the SEC approach, an action representation consists only of touchings (Ts) and untouchings (Us) between objects which are numbered consecutively by appearance. These points of touching and untouching (TUs, hereafter) were chosen as objective event boundary candidates in the current thesis to investigate event structure perception. Wörgötter et al. (2013) showed that computer vision using the SEC approach was able to distinguish between 30 different one-handed object manipulations typical of everyday life. Thus, this approach can represent object-directed actions, and we chose from these actions to build the stimulus material for the current work. To be precise, the TU sequence representation of an action provides its action category while the action kinematics further differentiate actions that share the same TU sequence.

To give an example, turning an object and pushing an object on a table both share the same TU sequence. Formulated in detail this is, the hand untouches the ground surface, the hand touches the object, the object is either turned or pushed, the hand detaches from the object and touches the ground surface to rest. The corresponding TU sequence reads U-T-U-T, while we simplified the SEC matrix and did not give the corresponding objects here. Actions that

can be represented by this specific TU sequence belong to the action category termed "rearrange" and differ from other action categories like, for instance, "break" that in turn includes actions like ripping-off and uncovering by picking and placing. Please note that the terminology of the action categories was adopted interdisciplinarily from the robotics perspective of Wörgötter et al. (2013) and the everyday understanding of the terms may differ from their use here.

For robots, this way of formalizing object-directed actions as sequences of relational changes between objects, hands and the ground, has proven highly useful to recognize human actions and to execute these actions itself (Aksoy et al., 2011). For cognitive neuroscience, the use of objective event boundaries and sparsely coded action categories that can be extracted directly from the stimuli offers promising opportunities to understand the neural processing underlying ongoing action observation and segmentation in the human brain.

1.3.2 From Robots and Infants

As mentioned above, robots can recognize and execute object manipulations using the SEC-based representation without prior object knowledge (Aksoy et al., 2011). For humans, TUs are salient and easily recognizable incidents, since touching is mostly accompanied by deceleration and untouching anticipates acceleration of our movements. They appear to be an ideal starting point for learning about action segments and action categories even before critical object expertise is built (for the early development of object knowledge see Hunnius & Bekkering, 2010). In line with this idea, it has been proposed that object-action association may develop earlier than object-word association (Eiteljoerge et al., 2019). Research points to the importance of developing event segmentation skills in early infancy to make sense of the world (Levine et al., 2019). Previous developmental work showed that infants at the age of two months detect structures inherent in the environment through statistical learning (Kirkham et al., 2002) and infants at the age of four months show a preference for biological motion patterns (Fox &

McDaniel, 1982)[3]. Accordingly, preverbal infants might identify TUs and use TU sequences to efficiently segment, encode and more easily recognize and predict everyday object manipulations they observe (cf. Wörgötter et al., 2020; Ziaeetabar et al., 2020). When they accrue greater experience with the world and access additional sources of information, they will be in the position to utilize this knowledge and yet TUs could remain relevant for processing in the brain. Although this remains purely speculative, it provides valuable new perspectives to consider.

[3] See Hunnius and Bekkering (2014) for a review of the development of action understanding abilities during childhood.

2 Research Questions and Objectives

As outlined in the preceding sections, there is a growing interest in understanding the neural processes underlying event structure perception and representation in action. While some light has already been shed, many questions remain unanswered. Furthermore, this field may profit tremendously from multidisciplinary perspectives and computer vision methods are well-suited to contribute.

The objective of this thesis was to investigate event structure perception and representation through objective (i.e., stimulus-derived) and subjective (i.e., observer-labeled) event boundaries as well as corresponding action categories. The employed object-directed actions involved either commonplace or dough items to modulate object-action associations. The neural responses of passive action observers were modeled with subjectively and objectively derived event boundaries and the neural response patterns were related to subjectively and objectively determined action categories. Based on the data obtained on segmentation and categorization, as well as their neural processing, the following research questions were addressed:

1. Are TUs as objective event boundaries a meaningful supplement or reference point to subjective event boundaries and how do they contribute to understanding neural event structure processing in object-directed action observation?

2. Do object-action associations, provided by the manipulated object, modulate action segmentation behavior and the neural processing at subjective and objective event boundaries?

3. Are TU action categories represented in neural processing patterns of object-directed actions and are they related to behavioral action classifications?

A series of two fMRI experiments with three tasks each was conducted to answer the targeted research questions. The two experiments were set up completely identically except for the stimulus material. Based on the SEC framework, short videos of simple object-directed hand actions were created and the degree to which the manipulated objects were associated with actions varied between experiments. The movements, in contrast, were kept constant, which allowed us to disentangle the effect of these two primary dimensions.

In the first experimental task, participants passively observed object-directed action videos during the MRI scan. The second task consisted of two behavioral sessions to determine the subjective event boundaries, namely a test and a retest session, where participants manually segmented the action videos by pressing a button whenever they thought an action step ended and a new began (cf. Newtson, 1973). Finally, the third one was a multi-arrangement task in which the participants spatially arranged the videos (represented by image triplets) according to their similarity (cf. Kriegeskorte & Mur, 2012) to derive action categories. In Experiment 1, the object-directed actions shown in the action videos were directed at commonplace items (e.g., a calculator, a piggy bank, or a cup) whereas in Experiment 2, manipulations of formed pieces of blue play dough were presented. Thus, as mentioned above, the manipulation remained consistent, though the manipulated item varied between experiments. Furthermore, the SEC algorithm was used to extract points of touching and untouching in the action videos which were subsequently employed as objective event boundaries. For clarity, “Experiment 1” and “Experiment 2” refer to the original empirical investigations conducted as part of this research project. The results of these experiments were subsequently published in three separate research articles, upon which this thesis is based, and are referred to here as Study I, Study II, and Study III.

Study I (Pomp et al., 2021) based on the first and second task of the first experiment (i.e., involving commonplace items). It investigated the relation of observer-labeled event boundaries

to stimulus-derived TUs in terms of temporal co-occurrence and neural processing. It was reasoned that if TUs are critical reference points for subjective action segmentation, they then show a systematic temporal relation to observer-labeled event boundaries or even match them. If both types of event boundaries coincide, we expected to replicate previously found brain activation patterns at event boundaries. In the case of the event boundaries being temporally distinguishable, we expected time-locked brain responses to also differentiate.

Study II (Pomp et al., 2024) used the data of the first and second task of both experiments. The segmentation and neural processing of commonplace item's manipulation and dough item's manipulation were compared elucidating the role of object-action associations. It was hypothesized that if the strength of object-action associations modulates subjective action segmentation, significant differences between segmentation behavior and neural processing in Experiment 1 and Experiment 2 will be found. Based on previous work, we hypothesized possible effects to be found in three regions of interest (ROIs): the anterior inferior parietal lobule (aIPL), the parahippocampal cortex (PHC), and a biological motion-sensitive area in the lateral temporo-occipital cortex (gratefully adopted from Hodgson et al., 2023). We assumed a knowledge-driven activation increase in the former two regions for commonplace items and a sensory-driven increase in the latter region for dough items.

Finally, Study III (Pomp et al., 2025) examined the neural representation of action categories across experiments. The subjectively determined action categories from the third task were used to model the participants' brain activity, alongside action categories derived from TU sequences and control models. We aimed to investigate which brain regions (if at all) reflect action categories as predicted by their TU sequence and therefore used representational similarity analyses (RSA; Kriegeskorte et al., 2008). In addition to brain-wide analyses, we closely examined the action observation network and therein we specifically focused on the left aIPL.

The guiding objective of the current thesis was to examine objective event boundaries and the advantages they can offer for event perception research. Especially as objective event boundaries can shift the focus from behavioral signatures of events and offer researchers an alternative way to explore how event structure perception occurs in the human brain during action observation. In addition, the value of the stimulus-driven action categorization for understanding neural action representation was to be determined. These boundaries and categories may be inherent to the stimulus, and the field would significantly benefit from identifying and understanding them and their further implications.

3 Research Articles

3.1 Study I: Touching Events Predict Human Action Segmentation in Brain and Behavior.

Running title:
3.1 Touchings Predict Human Action Segmentation

Jennifer Pomp, Nina Heins, Ima Trempler, Tomas Kulvicius, Minija Tamosiunaite, Falko Mecklenbrauck, Moritz F. Wurm, Florentin Wörgötter, & Ricarda I. Schubotz (2021)

NeuroImage, 243, 118534

https://doi.org/10.1016/j .neuroimage.2021.118534

Associated online data:

https://osf.io/jbwkq/?view_only=ae77638956974214a2faff6e674557a0 (OSF repository)

https://neurovault.org/collections/8736 (Brain activity maps on NeuroVault)

https://www.uni-muenster.de/IVV5PSY/AvicomSrv (Stimulus material on AVICOM)

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

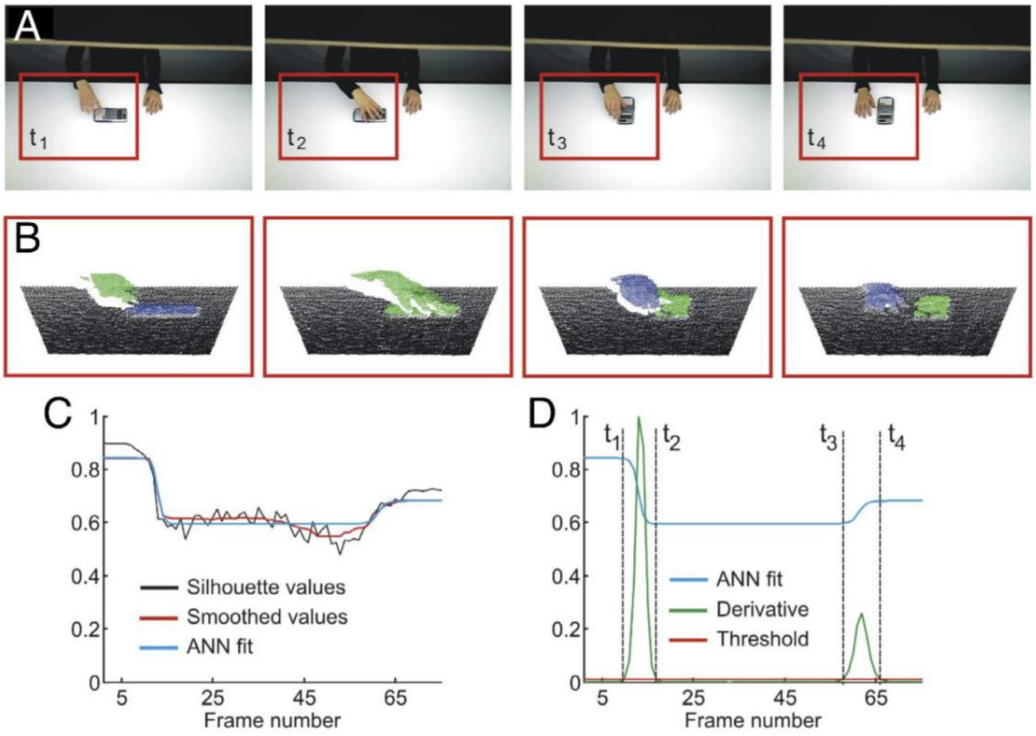

Fig. 1. Schema of the procedure for extracting the time points for touching and untouching events from an exemplary action, here "turning calculator". A) Point cloud extraction and preprocessing of RGB images. B) Clustering point clouds and calculating silhouette values. C) Curve fitting using artificial 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:

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

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

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 $d_{|t\text{-}t'|}$ 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.

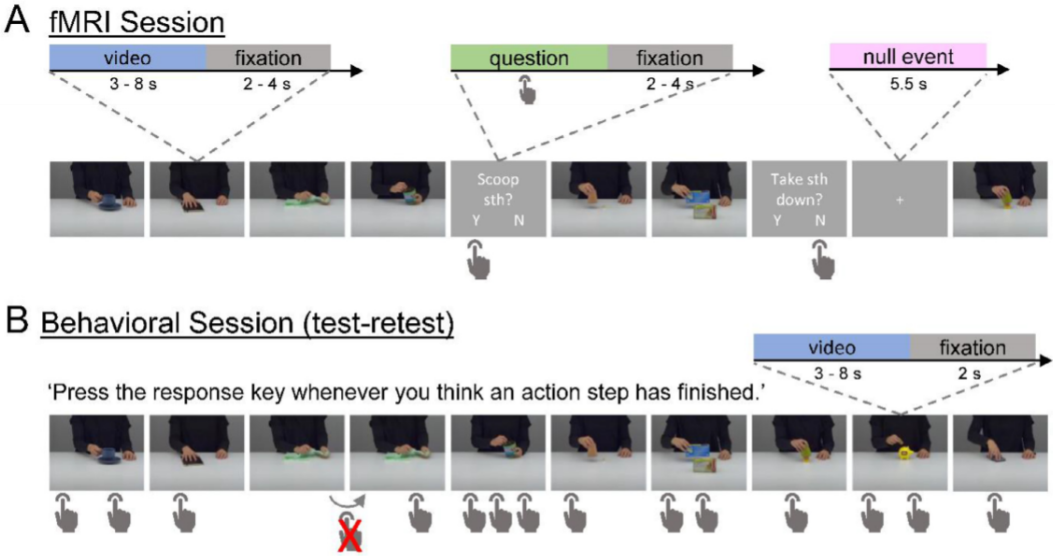

Fig. 2. Experimental design. (A) In the fMRI session, action trials and null trials were passively observed and question trials required participants to confirm or reject an action description with regard to the preceding action video. The question disappeared after button press. (B) In the two behavioral 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) ≈ 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

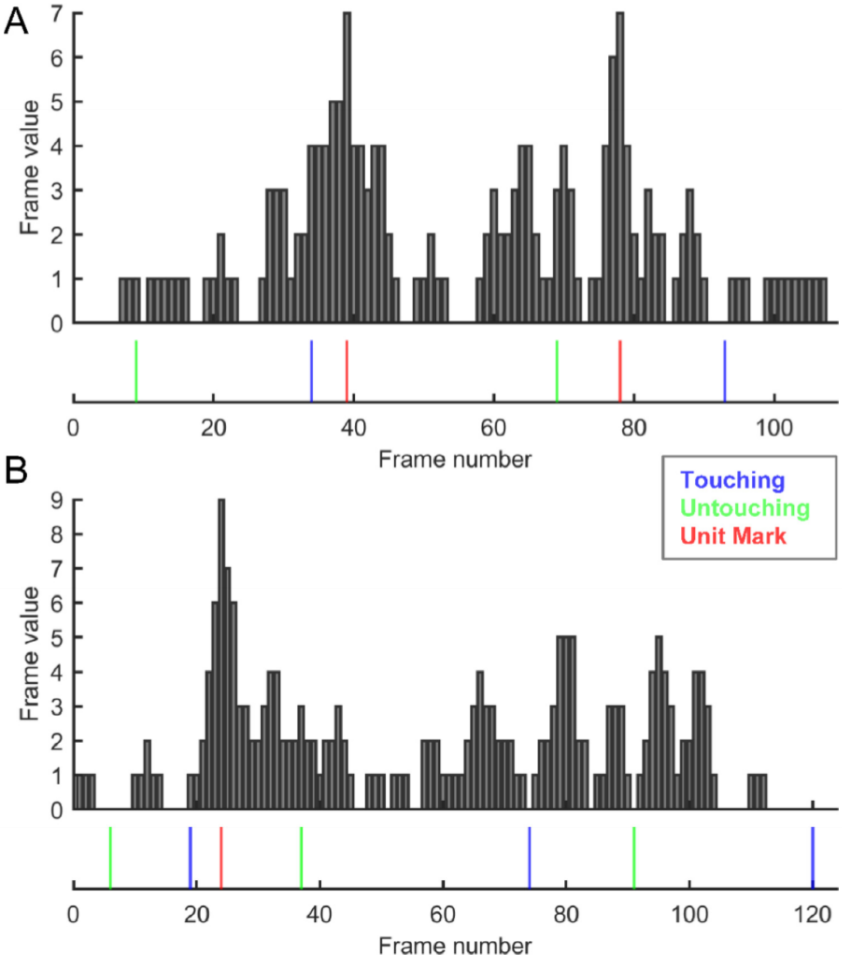

Fig. 3. Pooled unit marking responses of the group (n = 31) for two exemplary object manipulation videos: turning a bottle (A), putting a cup on top of a saucer (B). Maximum frame values were taken as group-consistent unit marks (Ms), as indicated in red on the lower x-axis. Respective touching (T) and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

rized 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

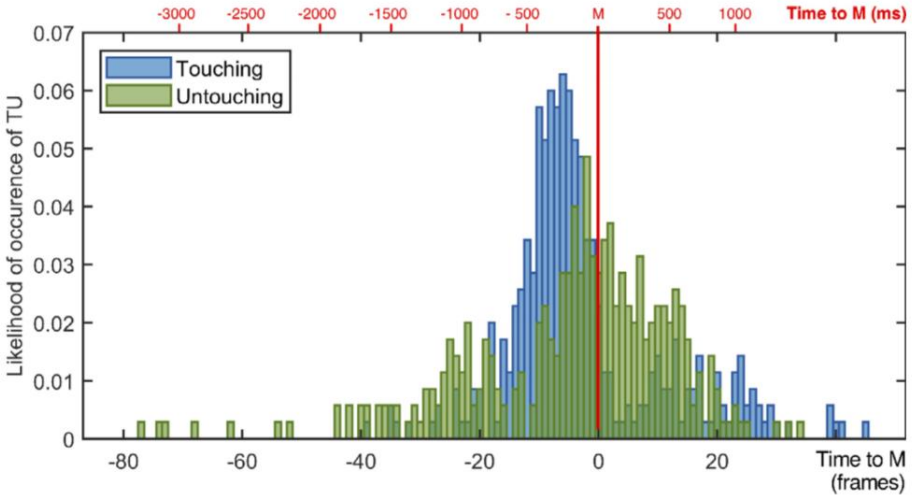

Fig. 4. Likelihood of occurrence of M-close touching events and M-close untouching events; the solid red line indicates the point in time where participants delivered a response for unit markings in the test-retest sessions (M), the lower x-axis shows the temporal distance of the events to M in frames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

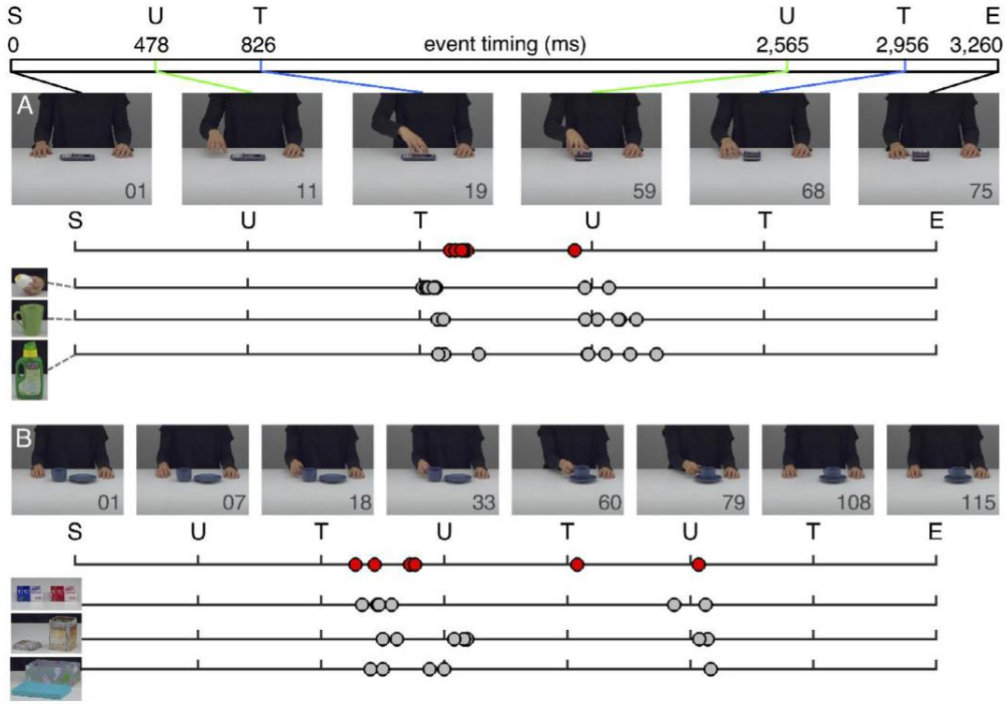

Fig. 5. Touching (T) and untouching (U) events as determined by computer vision for two exemplary object manipulation videos, and corresponding unit marks (M) delivered by participants. Single frame images are shown for all identified T and U events, with frame numbers given in the downright corner of 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.

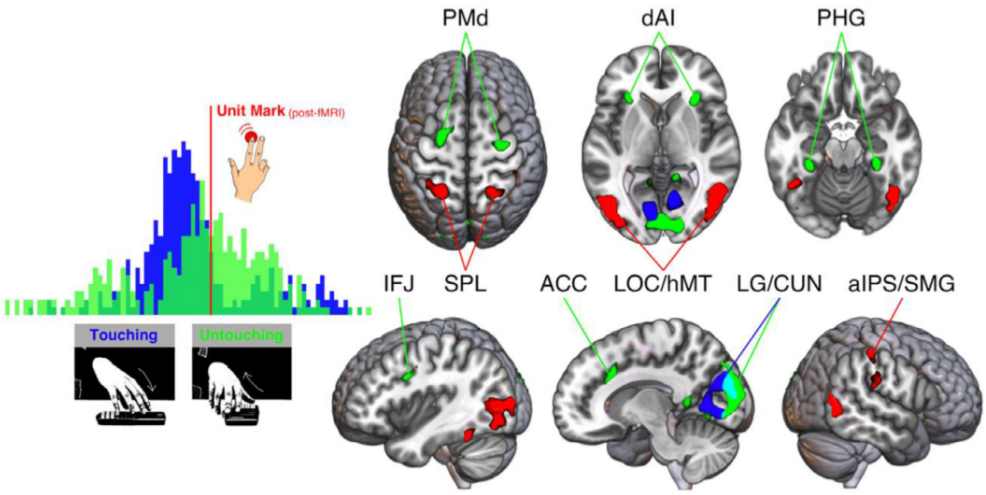

Fig. 6. Functional MRI activation at $p < .005$, peak-level FDR-corrected, for the main contrasts of post-fMRI human-determined unit marks (M>nM, red), objective touching events (T>nTU, blue) and objective untouching events (U>nTU, green). The overlap of the activation of touching and untouching in the 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 relation-

ship 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 occa-

sions 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 (AVICOM, 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.

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

Supplementary material

1. Tables

Table 1. List of individual object manipulations

Manipulation	Objects			
Turn	Calculator	Egg timer	Cup	Bottle of plant food
Pull	Notebook	Letter	Playing card	Flyer
Rip off	Garbage sack	Croissant	Masking tape	Notepad
Uncover	Dice cup, dice	Flowerpot, key	Newspaper, cellphone	Postcards
Take down	Two notepads	Two bowls	Two cups	Two toy blocks
Take away	Two shower gels	Two mandarins	Two marker pens	Two packets of tea
Put on top	Two packs of playing cards	Tea tin, lid	Container, lid	Saucer, cup
Put together	Two spice shakers	Two remote controls	Pen, pens	Piece of a puzzle, puzzle
Cut	Cake, knife	Paper, scissors	Fabric, scissors	Cardboard, carpet knife
Scoop	Container with ground coffee, measuring spoon	Cup of coffee, teaspoon	Sugar package, measuring spoon	Bowl of flour, hand
Hide	Plastic cup, marple	Brown envelope, letter	Egg cozy, egg	Folder, piece of paper
Put into	Hard disk, case	Wallet, bank bill	Cup of tea, sugar cube	Piggybank, coin

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

Table 2. Analysis of sequential motifs embedding M

Triplet	M embedded in triplet			General occurrence per triplet	
	Observed number	Likelihood	Expected number	Observed number	Likelihood
U-T-U	176***	0.503	129.4	411	0.412
T-U-T	79***	0.226	133.1	423	0.424
U-T-T	28	0.080	22.9	73	0.073
T-U-U	14**	0.040	5.7	18	0.018
T-T-U	10	0.028	15.4	49	0.049
T-T-T	6	0.017	1.9	6	0.006
U-U-T	1	0.003	5.7	18	0.018
U-U-U	0	0.000	-	0	0.000

Note. T = Touching event, U = Untouching event, M = Unit mark. Asterisks indicate whether the observed number of triplets embedding M significantly differed from the expected number with *** = $p < .001$ and ** = $p < .005$.

2

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

Table 3. Video content analysis

Action phase	M placed in action phase		General occurrence	
	Observed number	Expected number	Observed number	Likelihood
Object manipulation	74***	39.64	128	0.14
Start of object transport	82***	45.83	148	0.16
Hand transport	44***	72.15	233	0.26
Object transport	39*	60.69	196	0.22
Tool transport	4	3.72	12	0.01
End of object transport	26**	49.54	160	0.18
end of manipulation with a tool	10	7.43	24	0.03

Note. M = Unit mark. Asterisks indicate whether the observed number significantly differed from the expected number with *** = $p < .001$, ** = $p < .005$, and * = $p < .05$.

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

Table 4. Maxima of activation from the direct contrasts of the second-level whole-brain analyses at p <.005 peak-level FDR-corrected.

Macroanatomical location	H	Cluster Extent	*t*-value	MNI Coordinates x	y	z
M > T ∩ M > U						
Anterior inferior parietal sulcus	L	266	7.32	-57	-22	29
Superior parietal lobule	L		5.34	-33	-46	62
Anterior inferior parietal sulcus/ Supramarginal gyrus	R	209	7.97	60	-16	29
Ventral premotor cortex / inferior frontal gyrus (BA 6/44)	R	52	6.32	54	11	14
Fusiform gyrus	L	40	6.30	-45	-52	-19
	R	32	5.05	45	-52	-13
Lateral occipital cortex / human motion area	L	69	5.85	-42	-67	5
Mid-insula	L	23	6.21	-39	-4	14
	R	23	5.34	42	-1	11
T > M						
Cuneus	L	20	6.00	-6	-82	26
	R	59	7.63	12	-76	23
Lingual gyrus	L	42	7.60	-9	-79	-1
	R	11	5.47	12	-79	-4
U > M						
Cuneus	L	963	9.65	-9	-97	14
	R		7.82	9	-94	14
Parieto-occipital fissure	R	114	5.86	21	-58	29
Retrosplenial cortex	R		4.90	9	-46	5
Parieto-occipital fissure	L	61	5.18	-18	-58	23
Parahippocampal gyrus	R	38	5.71	24	-46	-4
Dorsal premotor cortex	L	123	6.76	-21	2	56

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

	R	134	5.55	33	-13	59
			5.54	21	-4	56
Anterior superior frontal sulcus	L	35	5.34	-30	44	17
	R	48	5.13	27	38	26
Dorsal anterior insula	L	47	6.59	-27	26	-1
	R	62	5.19	27	20	-1
Anterior cingulate cortex	R	82	5.95	12	20	32

Note. H = Hemisphere, MNI = Montreal Neurological Institute, L = Left, R = Right, M = Unit mark, T = touching event, U = untouching event.

Table 5. Maxima of activation from the main contrasts of the second-level whole-brain analyses at p <.005 peak-level FDR-corrected of the parametric GLM.

Macroanatomical location	H	Cluster Extent	*t*-value	MNI Coordinates x	y	z
Mp						
Lateral occipital cortex / human motion area	L	4760	10.95	-51	-73	5
	R		10.22	48	-70	-7
Fusiform gyrus	L		8.81	-39	-49	-19
Superior parietal lobule	L		6.17	-18	-52	71
	R		5.94	21	-55	68
Inferior parietal lobule	L		6.26	-51	-37	53
	R		7.46	36	-52	59
Angular gyrus	L		5.57	-39	-61	47

5

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

Dorsal premotor cortex	L	151	6.92	-48	11	44
	R	250	6.72	51	5	50
Inferior frontal gyrus pars triangularis	L	272	7.79	-57	17	20
Mid-Insula	L		4.49	-45	-1	5
	R	16	4.40	42	-1	5
Inferior frontal gyrus pars opercularis	R	97	5.50	57	11	11
Inferior frontal gyrus pars triangularis	R		4.87	57	32	17
Supramarginal gyrus	R	333	6.26	63	-28	38
Thalamus	R	34	6.29	15	-28	2
TU > nTU						
Cuneus	L	1697	8.52	-9	-97	17
	R		8.43	15	-91	29
Lingual gyrus	L		7.87	-6	-79	-1
	R		6.16	12	-79	-4
Fusiform gyrus	R		6.16	33	-52	-7
Parahippocampal gyrus	R		4.71	30	-37	-16
T > nTU						
Lingual gyrus	L	1136	8.75	-9	-76	-1
	R		7.59	12	-76	-4
Cuneus	L		7.84	-9	-88	23
	R		7.17	9	-79	26
U > nTU						
Lingual gyrus	L	1462	10.05	-24	-73	-4

	R		9.12	33	-52	-7
Cuneus	L		8.71	-9	-100	17
	R		8.28	15	-94	29
Parahippocampal gyrus	L		6.39	-30	-34	-16
	R		5.83	30	-31	-16
Parieto-occipital fissure	R		5.82	21	-55	23
	L	34	5.03	-15	-67	29
Dorsal premotor cortex	L	203	7.32	-24	2	53
	R	175	6.58	24	2	50
Anterior superior frontal sulcus	R	15	4.86	27	35	29
Inferior frontal junction	L	27	5.24	-36	5	29
Dorsal anterior insula	L	27	5.78	-27	23	-1
	R	69	6.06	30	23	5
Dorsal anterior cingulate cortex	R	32	5.53	12	20	32

Note. GLM = General linear model, H = Hemisphere, MNI = Montreal Neurological Institute, L = Left, R = Right, Mp = parametric unit mark, T = touching event, U = untouching event, nTU = non-touching/untouching event.

J. Pomp et al. https://doi.org/10.1016/j.neuroimage.2021.118534

2. Figures

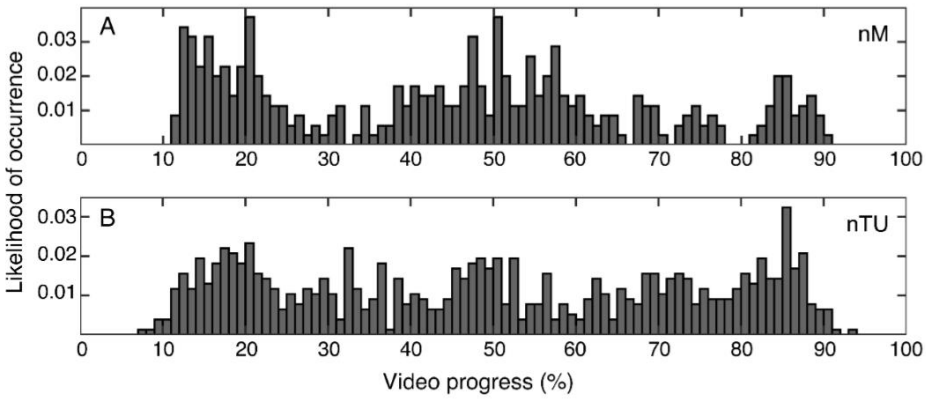

Figure 1. Likelihood of occurrence of non-critical events during the video, nM = non-unit mark, nTU = non-(un)touching event. A) Likelihood for non-unit marks. B) Likelihood for non-(un)touching events.

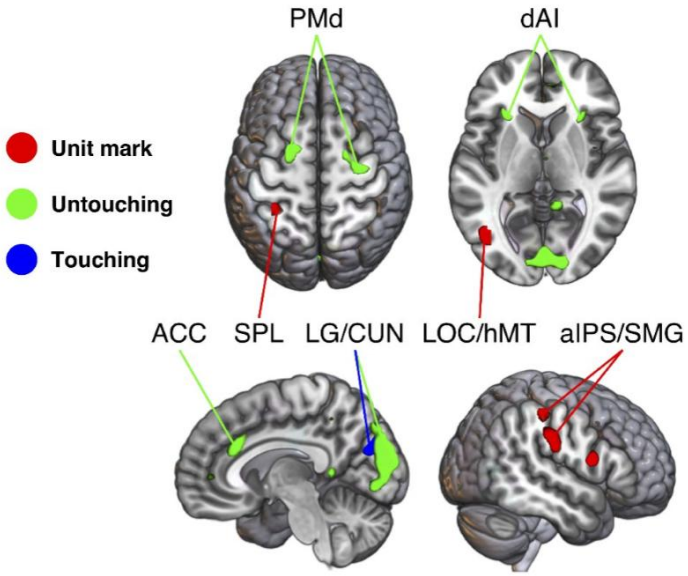

Figure 2. Functional MRI activation at $p < .005$, peak-level FDR-corrected, for the effects based on direct contrasts between experimental conditions: the conjunction contrast of

post-fMRI human-determined unit marks versus untouching and untouching [(M > T) ∩ (M > U), red], objective untouching events (U>M, green) and objective touching events (T>M, blue). Peak coordinates are given in Supplementary Table 4. PMd = dorsal premotor cortex, dAI = dorsal anterior insula, 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

3.2 Study II: Action Segmentation in the Brain: The Role of Object–Action Associations.

Running title:

3.2 Object-Action Associations in Action Segmentation

Jennifer Pomp, Annika Garlichs, Tomas Kulvicius, Minija Tamosiunaite, Moritz F. Wurm, Anoushiravan Zahedi, Florentin Wörgötter, & Ricarda I. Schubotz (2024)

Journal of Cognitive Neuroscience, 36:9, 1784-1806

https://doi.org/10.1162/jocn_a_02210

Associated online data:

DOI10.17605/OSF.IO/MGQSF (OSF repository)

https://neurovault.org/collections/16065 (Brain activity maps on NeuroVault)

https://www.uni-muenster.de/IVV5PSY/AvicomSrv (Stimulus material on AVICOM)

Action Segmentation in the Brain: The Role of Object–Action Associations

Jennifer Pomp[1,2], Annika Garlichs[1], Tomas Kulvicius[4], Minija Tamosiunaite[3,5], Moritz F. Wurm[6], Anoushiravan Zahedi[1,2], Florentin Wörgötter[3], and Ricarda I. Schubotz[1,2]

Abstract

■ Motion information has been argued to be central to the subjective segmentation of observed actions. Concerning object-directed actions, object-associated action information might as well inform efficient action segmentation and prediction. The present study compared the segmentation and neural processing of object manipulations and equivalent dough ball manipulations to elucidate the effect of object–action associations. Behavioral data corroborated that objective relational changes in the form of (un-)touchings of objects, hand, and ground represent meaningful anchor points in subjective action segmentation rendering them objective marks of meaningful event boundaries. As expected, segmentation behavior became even more systematic for the weakly informative dough. fMRI data were modeled by critical subjective, and computer-vision-derived objective event boundaries. Whole-brain as well as planned ROI analyses showed that object information had significant effects on how the brain processes these boundaries. This was especially pronounced at untouchings, that is, events that announced the beginning of the upcoming action and might be the point where competing predictions are aligned with perceptual input to update the current action model. As expected, weak object–action associations at untouching events were accompanied by increased biological motion processing, whereas strong object–action associations came with an increased contextual associative information processing, as indicated by increased parahippocampal activity. Interestingly, anterior inferior parietal lobule activity increased for weak object–action associations at untouching events, presumably because of an unrestricted number of candidate actions for dough manipulation. Our findings offer new insights into the significance of objects for the segmentation of action. ■

INTRODUCTION

Everyday actions consist of smoothly concatenated action steps. The segmental structure of actions is reflected in the way that we teach, learn, and execute actions ourselves (Braun, Mehring, & Wolpert, 2010), and also in how we perceive actions performed by others (Newtson, Hairfield, Bloomingdale, & Cutino, 1987). Behavioral studies in children (Buchsbaum, Griffiths, Plunkett, Gopnik, & Baldwin, 2015; Baldwin, Baird, Saylor, & Clark, 2001) and adults (Hard, Recchia, & Tversky, 2011; Newtson & Engquist, 1976) show that action segmentation arises spontaneously (see also Zacks, Speer, Swallow, Braver, & Reynolds, 2007) and helps us process and remember dynamic events efficiently (Kurby & Zacks, 2018; Zacks & Swallow, 2007).

To measure subjective segmentation behavior, researchers ask participants to indicate when they perceive event boundaries, that is, those points in time when one action segment ends and the next begins (Newtson, 1973). This procedure has been shown to yield intra-individually consistent action segments (for a review: Sargent, Zacks, & Bailey, 2015), but the question remains which stimulus properties drive the segmentation behavior. A number of studies have specifically addressed the role of motion as a cue for updating action models at event boundaries (Zacks, Kumar, Abrams, & Mehta, 2009; Hard, Tversky, & Lang, 2006; Newtson, Engquist, & Bois, 1977). Typical measures to quantify motion include binary time interval coding for separate movement types (Hard et al., 2006) or motion tracking through speed and acceleration of hands and head (Zacks et al., 2009). Correspondingly, the activity of the motion-selective middle temporal visual area (MT/V5) was found to increase during the perception of event boundaries in actions (Schubotz, Korb, Schiffer, Stadler, & von Cramon, 2012; Speer, Swallow, & Zacks, 2003; Zacks et al., 2001), pointing to change in motion as an efficient cue that announces event boundaries and triggers updating processes in frontal networks.

However, having a life-long experience with manipulable objects, the movements one expects when observing object-directed actions certainly also depend on the involved object and might influence spatial attention and processing. Objects are an important source of information that individuals use to understand an observed action because we have learned how to act with or on an object and thereby build object–action associations (Borghi,

[1]University of Münster, [2]Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, [3]University of Göttingen, [4]University Medical Center Göttingen, [5]Vytautas Magnus University, Kaunas, Lithuania, [6]University of Trento

Journal of Cognitive Neuroscience 36:9, pp. 1784–1806
https://doi.org/10.1162/jocn_a_02210

2021; Zhao, 2019). In a former study (Schubotz, Wurm, Wittmann, & von Cramon, 2014), we built on the idea that objects are reminiscent of actions often performed with them. For instance, the combination of a knife and an apple remind us of peeling the apple or cutting it. Findings confirmed that the BOLD response in action-related inferior parietal and posterior temporal areas varied with the number of object-implicated actions. This impact of objects has been shown to influence the processing of observed action, even when these objects are not actually used (El-Sourani, Trempler, Wurm, Fink, & Schubotz, 2019; El-Sourani, Wurm, Trempler, Fink, & Schubotz, 2018; Hrkać, Wurm, Kühn, & Schubotz, 2015). However, because action segmentation appears to be highly dependent on movement-related information and may develop in early infant action observation when functional or semantic knowledge about objects is still rudimentary, object information may not be essential for action segmentation. One may ask how action structures are processed before having experience-based knowledge of object-associated actions, for instance, when encountering actions with novel objects, which is common in young infancy (cf. Hunnius & Bekkering, 2010).

In the present study, we aimed to investigate the effect of object–action knowledge on action segmentation and underlying brain processes. We built on a previous study (Pomp et al., 2021), which examined action segmentation in everyday object manipulations. To this end, we recreated the movies of the object manipulation actions, but this time using formed pieces of play dough as objects. This replacement of common objects by formed dough minimized object–action associations, that is, individuals did not strongly associate the formed dough with specific actions (except for kneading, if at all). The actions themselves were kept as similar as possible to the actions performed on the everyday objects to balance the movement patterns between the current and the previous study. After a passive action observation session in the MRI scanner, individual behavioral action segmentations of these actions were gained using the unit marking procedure (Newtson, 1973). Although subjective reports are important and can be informative, we do not necessarily have explicit access to all event boundaries that our brain registers and exploits to make sense of the world. Moreover, subjective reports may be focused on behaviorally relevant events and have been shown to be highly dependent on the exact task, for example, with regard to the instruction of detecting "meaningful" boundaries or selecting a specific "fine" or "coarse" grain of the segmentation (Zacks et al., 2007). Manual action segmentation is therefore a possible, but not necessarily a reliable, approximation for the way in which the brain segments events.

An exciting complement to research into action segmentation is therefore a more objectifiable stimulus-based approach to action segmentation (Pomp et al., 2021). We extracted objective stimulus characteristics based on the notion of *semantic event chains* (Wörgötter et al., 2013; Aksoy et al., 2011). In an object-directed action, this approach describes actions as a sequence of relational changes in the form of *touchings* (T) and *untouchings* (U) of objects, hands, and ground (TUs, hereafter). For instance, when a hand grasps an object, motion velocity usually reaches zero when the hand and object touch. In case of a subsequent object transport, the object then untouches the ground, and velocity increases again until it decreases before the object touches its destination. In case of a subsequent object manipulation, for example, turning, velocity increases while the object is turned and decreases before the hand untouches the object after manipulation. Thus, the binary coding of touching relations (touch, untouch) between each pair of objects, hands, and ground in an action scene can be used to describe the course of action without the need to analyze velocity and trajectory patterns and was used in the current study to model brain activity. Note that the above-explained underlying computer-vision algorithm that we used is model-free and stimulus-driven (Aksoy et al., 2011). Therefore, it does not require functional or semantic knowledge about objects (or hands or ground), which might imitate the simple model of early infant action observation. The use of objective event boundaries, which can be extracted directly from the stimulus material, offers promising opportunities to understand the neural processes underlying ongoing action segmentation.

Using the touching–untouching approach in the present study, we examined the impact of object–action knowledge on action segmentation and underlying brain processes. If object–action associations play a role in action segmentation, we expected significant differences between our previous study on object manipulation and our current study on dough manipulation in terms of segmentation behavior and time-point-specific brain activity. To this end, we compared the neural processing of object and dough videos at different types of event boundaries, including group-consistent behavioral segmentations (unit marks, Ms hereafter) and objective TU events as relevant points in time. We refer to the boundaries assessed by the participants as unit marks (conceptually based on the unit marking procedure) and not as event boundaries, as we assume that they are only one type of event boundary of interest. For object manipulations, TU events were found to be meaningful anchor points for action segmentation behavior (Pomp et al., 2021), and we expected TUs to gain even more importance when object–action associations are weak. Specifically, we expected participants' action segmentation behavior to be even more dependent on TU events, that is, temporally less spread and closer to TUs. We refer to the temporal relation between participant-judged event boundaries and TU events as being *systematic* if their occurrence coincided more than randomly often, which we examined on single subject and group level. For object manipulations, this systematic relation had been shown (Pomp et al., 2021) and we expected that this systematicity in behavior would increase for

dough manipulations. Thus, we expected that participant-judged event boundaries would reliably coincide with TU events, but not necessarily vice versa.

With regard to brain activity, we examined at which of the critical time points T, U, and M activity would differ between object and dough manipulation in one of three ROIs derived from previous findings: the anterior inferior parietal lobule (aIPL), the parahippocampal cortex (PHC), and the biological motion-sensitive area (BMA, hereafter) in the lateral temporo-occipital cortex. Concerning the first ROI, as mentioned above, Schubotz and colleagues (2014) showed inferior parietal regions' activity to vary with the number of object-implicated actions at the mere sight of the object, independent of its usage. This activity was located in aIPL, and therefore, we expected increased aIPL activation for actions performed on objects versus dough pieces. The aIPL, as part of the ventrodorsal visual processing route (Binkofski & Buxbaum, 2013), is engaged in the representation of pragmatic object properties (Bosch et al., 2023) and hand–object interactions (Pelgrims, Olivier, & Andres, 2011; Vingerhoets, 2008) when we perform, plan, or observe object manipulations. Correspondingly, aIPL is known to be an important anatomic substrate underlying ideomotor apraxia (O'Neal et al., 2021), and it has been suggested to resolve competition between possible actions (Watson & Buxbaum, 2015). Concerning the second ROI, as for aIPL, we hypothesized an increased PHC activation for actions performed on objects versus actions performed on dough. The PHC is generally involved in processing contextual associations (Li, Lu, & Zhong, 2016; Aminoff, Kveraga, & Bar, 2013; Bar, Aminoff, & Schacter, 2008), which is the principal element underlying many cognitive processes, including spatial processing in scenes and episodic memory. In previous studies, we found PHC activity to specifically increase at action boundaries, possibly signaling the memory-driven updating of expectations of the next action associated with the object (Pomp et al., 2021; Schubotz et al., 2012). We here expected that familiar objects would trigger more contextual action associations than formed pieces of play dough accompanied by higher PHC activity. Finally, regarding the third ROI, we expected motion information to gain importance for play dough compared with object videos, which we hypothesized to detect in BMA. We reasoned that detailed motion analysis might be less critical when objects provide clues about which actions are about to be performed, whereas detailed motion analysis might be especially important, when pieces of dough are manipulated, to constrain the observer's predictions efficiently.

METHODS

For the current study, we used the experimental design of a previous study (Pomp et al., 2021), employed new videos, and tested a new group of participants comparable in size. The current study was kept as similar as possible to the previous one to allow direct statistical comparisons. This includes that the participants were recruited through the same channels, the study took place at the same institute, participants were scanned in the same MRI scanner, behavioral sessions were in the same laboratory rooms, and all sessions followed the exact same experimental protocols with similar equipment and materials (except for the stimulus videos). The results of the previously published study will not be shown here again, but only new analyses relating to statistical between-studies comparisons. Regarding brain activity contrasts, only interaction effects are reported, to make the results resistant to any differences between groups. With regard to the interpretation of direct comparisons between the two studies, we statistically compared the sample characteristics to rule out that differences between the samples could account for differences observed between the video types. We used the demographic details on age, sex, and profession, as well as participants' answers to the short surveys about their physical and mental condition, and experimental task features that concluded each of the separate sessions (for details on the survey, see section Experimental Procedure) to predict the participant's affiliation to either study. In separate analyses for the continuous, ordinal, and binary data types, no significant differences between groups were found using Bayesian modeling. To be precise, these analyses yielded support for the null hypothesis in all but one case, where the evidence ratio was inconclusive—giving neither evidence for the null nor for the alternative hypothesis. We uploaded the corresponding data, the R script of the analyses, and the results to the Open Science Framework (OSF) repository (DOI 10.17605/OSF.IO/MGQSF).

Participants

Thirty-three right-handed participants (M_{age} = 23.03 years, SD = 3.06 years, age range = 18–29 years, 28 women, 5 men) took part in this study. This sample size was based on previous work (Pomp et al., 2021) that showed robust results with a similar sample size. All participants reported intact color perception, and none of the participants reported any history of neurological or psychiatric disorders. The participants had not taken part in related precursor studies. In the course of the experiment, it became apparent that one participant had not understood the instructions of the behavioral segmentation task correctly; hence, this participant's data set was excluded from the behavioral model construction but was included in the fMRI data set (as the fMRI session was before and independent of the behavioral categorization task). Therefore, in the behavioral analysis, the data of 32 participants (27 women, 5 men) aged between 18 and 29 years (M = 22.88 years, SD = 3.13 years) were considered. Participants gave written consent to voluntarily participate in the experiment and were self-reportedly suitable for fMRI measures. They either received course credits or were paid for their participation. The current study is in

accordance to the Declaration of Helsinki and was approved by the local ethics committee of the Faculty of Psychology at the University of Münster (Germany).

Stimulus Material

The transitive actions employed in this study were designed based on the Semantic Event Chain (SEC) framework described by Wörgötter and colleagues (2013). Only transitive actions involving one active hand and one or two objects are included in this framework whereof 12 actions were selected for the current study that belonged to six action categories. The 12 selected actions were: turn, pull, rip off, uncover, take down, take away, out on top, put together, cut, scoop, hide, and put into. The execution of these transitive actions was recorded using an industrial camera (BASLER acA 1300-75gc) 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.) recording color as well as depth images. The video material presented in this study showed an actress from the front (BASLER camera) up to the shoulders performing the action with formed pieces of blue play dough on a white table. The ASUS Xtion Live recorded the actions from above, and its recordings were utilized for SEC time point extraction. For each object manipulation, 24–25 unique video takes were chosen for the final stimulus set (to account for the natural variation usually observed in human action performances), 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 five frames after the hand lies back on the table with a video duration ranging from 68 frames to 165 frames ($M = 112.35, SD = 18.13$), that is, 2957 msec to 7174 msec ($M = 4885, SD = 788$). To increase the perceptual variability, all videos were vertically mirrored so that actions seemed to be performed by the left hand. Each participant saw 50% of the actions mirrored.

Adopted from our previous study (Pomp et al., 2021), the stimulus sequence was designed as a second-level counterbalanced De Bruijn sequence with seven conditions (six action categories + null condition) created using the De Bruijn cycle generator (Aguirre, Mattar, & Magis-Weinberg, 2011). Subsequently, condition labels of the six action categories were permuted to create 20 different stimulus lists. Per list, half of the stimuli were shown mirrored, and a second list contained the complement of these, which gave 40 different stimulus lists in total. For the second and third experimental sessions, the start of the individual stimulus sequence was shifted by one third and two thirds, respectively, to prevent recognition of the stimulus sequence as well as to prevent time-dependent effects. For the fMRI session, the stimulus sequence was subdivided into seven runs, and at the start of each run, the last two videos of the preceding run were repeated and then discarded from analyses to presume a continuous stimulus sequence (the first run started with the last two videos of the last run).

Video Segmentation and SEC Determination

As previously described (Pomp et al., 2021), we used an automated extraction of time points of TU events. Extracting these TU events automatically had the advantage that human bias could be avoided in the objective segmentation process. A flow diagram for the automated extraction of time points at which touching/untouching relations between object pairs change is shown in Figure 1. Here, we used the frame number to define the time points. The input to the algorithm is a sequence of RGB-D frames f_i ($i = 1 \ldots n$, n is the number of frames), and the output is a sequence of time events t_i ($i = 1 \ldots$ 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.

Point Cloud Extraction and Preprocessing

Point clouds for each frame f_i were generated from depth images, which were acquired using ASUS Xtion Live sensor. ROI on the left side of the frame was cut as shown in Figure 1, because always only one hand was involved in the analyzed actions. Furthermore, point clouds were subsampled by a factor of four to reduce the number of points, this way speeding up the clustering procedure. Before clustering, ground plain subtraction was performed. Ground plain subtraction, that is, removing points

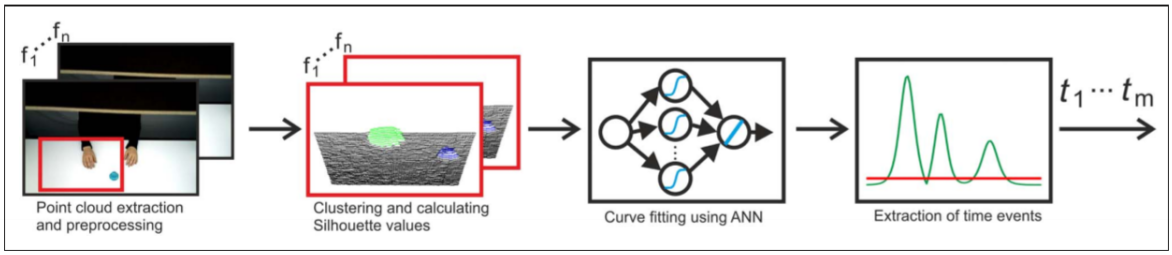

Figure 1. Flow diagram for the automated extraction of time points of TU events (see Methods section for details). ANN = artificial neural network.

corresponding to the table, was done as follows. First, we fitted a flat 2-D surface and then removed all points from the 3-D point cloud data, which were above the fitted plane, that is, we first removed points $p_i = \{x_i, y_i, z_i\}$, if $z_i - Z1_i > th1$, where $Z1_i = P1(x_i, y_i)$ are corresponding points of the fitted plane $P1$, and $th1 = 0.02$ is the manually set threshold. Afterward, we fitted the plane one more time to the remaining background points $bg\_p_i = \{bg\_x_i, bg\_y_i, bg\_z_i\}$ and we removed points that were below the fitted plane (see black points in Figure 2A, bottom row), that is, $p_i = \{x_i, y_i, z_i\}$, if $z_i - Z2_i < th2$, where $Z2_i = P2(bg\_x_i, bg\_y_i)$ are corresponding points of the fitted plane $P2$, and $th2 = 0.01$ is the manually set threshold. The removed points p_i were not included to further cluster analysis. Thus, for the clustering step, we only used point clouds of the hand and objects.

Clustering and Calculation of Silhouette Scores

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

$$S(f_i) = max(S_k), (k = 1 \ldots K), \text{with} \quad (1)$$

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

where $D_{within}(j)$ is the average distance from the jth point to the other points in its own cluster, and $D_{between}(j, l)$ is the average distance from the jth 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 score $S(f_i)$ decreases, whereas it increases when clusters are moving apart (see Figure 2B).

Fitting of Silhouette Curve Using Artificial Neural Network

The time points of TU events can be extracted from the Silhouette curve; however, Silhouette scores are noisy because of noise present in the point cloud data obtained from the RGB-D sensor. Thus, we first filtered the Silhouette scores

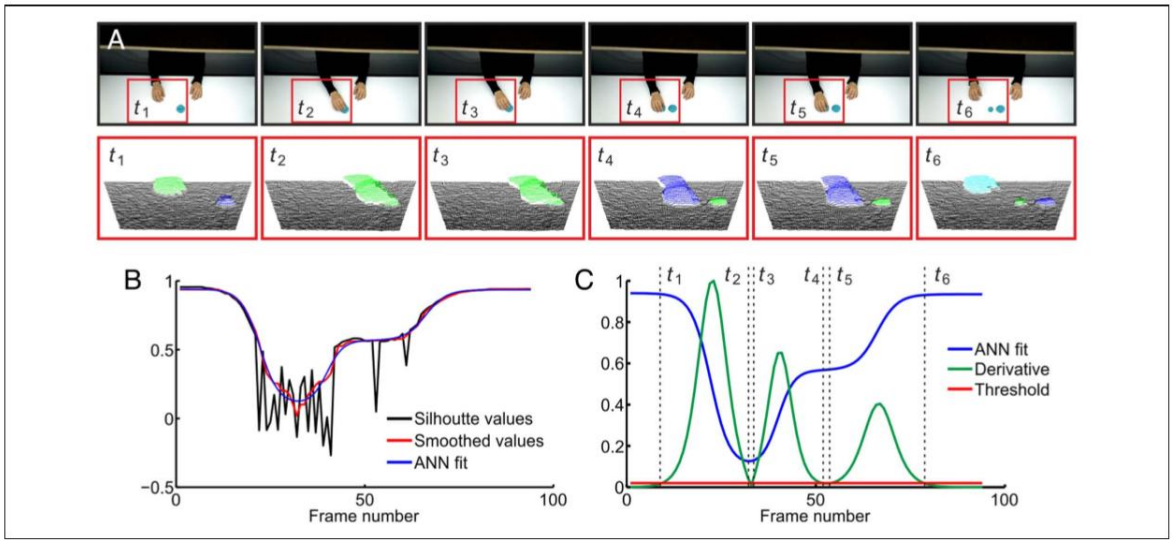

Figure 2. Schema of the procedure for automatically extracting the time points for touching and untouching events from an exemplary action, here "take down." (A) RGB images (top) from the above-scene installed ASUS Xtion Live RGB-D sensor and corresponding clustered point clouds (bottom). Clustered point clouds (objects) are color-coded and when two objects touch, they become one cluster with a shared color. When these objects untouch, the point clouds separate and one cloud changes to an individual color. (B) Raw silhouette values (black), smoothed silhouette values using a median filter (red), and fitted silhouette curve using an artificial neural network (ANN; blue). (C) Derivative of the ANN fit (green) and obtained time points of TU events after thresholding: t_1 = hand detaches from the table (i.e., first untouching); t_2 = hand touches the upper play dough object (i.e., first touching); t_3 = hand lifts the upper play dough object from the bottom play dough object (i.e., second untouching); t_4 = hand places the play dough object on the table (i.e., second touching); t_5 = hand detaches from the play dough object (i.e., third untouching); and t_6 = hand touches the table (i.e., third touching). Thus, in this example, a U-T-U-T-U-T event sequence is extracted. A demo source code of automated extraction that corresponds to the shown example can be downloaded from the OSF repository (DOI 10.17605/OSF.IO/MGQSF).

$S(f_i)$ 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 that 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 Figure 2B), which corresponded to changes in cluster configuration, that is, 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 Figure 2 for a "take down" action, we have six TU events (hand lifts up from the table, hand touches upper play dough object, hand lifts the upper play dough object from the lower play dough object, hand places the play dough object on the table, hand leaves the play dough object, 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 Figure 2C (see t_1, t_2; t_3, t_4; and t_5, t_6). The number of neurons h in the hidden layer was set based on the number of TU events m, that is, h = round(m/2). In this case, we used three neurons in the hidden layer. The network was fitted 10 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.

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 started with some initial threshold value $TH_{ini} = 0.01$ and increased it by 0.005 until the predefined number of TU time points was obtained. The time points were extracted at the frame numbers where the derivative of the $S_{ANN}(f_i)$ crossed the threshold value TH (see Figure 2C).

Whenever the algorithm misinterpreted the scene, which gave an error message, the extracted time points were checked against manual TU segmentation results and time points. Deviation from human TU segmentation, on average, was 4.14 frames ($SD = 3.42$), and in 93.02% of the cases, deviation was less than 10 frames (i.e., approx. mean value $+2 \times SD$). Thus, we corrected outliers in 6.98% of the cases, where TU event segmentation differences were larger than nine frames, by setting values of automated segmentation to corresponding values of human TU segmentation. The framework was implemented using MATLAB (https://www.mathworks.com) 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 corresponding non-events (nTU) to be maximally far away from an event.

Experimental Procedure

Congruent with our previous study (Pomp et al., 2021), participants completed three sessions. The MRI session was, on average, 4 days (range = 3–6) before the behavioral test–retest sessions, which were, on average, 14 days apart from one another (range = 14–18). In the first session, participants paid attention to the action videos while being in the MRI scanner. Action videos were back-projected onto a screen and displayed centrally with a screen resolution of 640 × 512 pixels by Presentation 20.3 (Neurobehavioral Systems Inc.). Participants viewed the screen binocularly through a mirror above the head coil. Attention-capturing questions followed 14% of the videos, asking whether an action description was appropriate for the preceding action video (see Figure 3A for the experimental trial design). Participants responded by pressing one of the two response keys with their right index and middle finger. 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 and 120 min including consent forms, instructions, preparation, scanning, and a short survey at the end.

The second experimental session comprised the unit marking task (Newtson, 1973). Participants saw the same videos as in the first session. Stimuli were presented on a 23-in. monitor by Presentation 18.1 (Neurobehavioral Systems Inc.), and participants were instructed to press a button with their right index finger whenever they think an action step is finished, that is, an event boundary occurred. Training trials were offered at the beginning, and two self-paced breaks were provided after one third and two thirds of the trials. This task took approximately 45 min. See Figure 3B for the experimental trial design. In the third session, this task was repeated to retest the unit marking behavior.

At the end of each of the three sessions, participants filled in a two-paged survey about their current state (mood, subjective health, tiredness before and after the task); the amount of sleep in the last night and whether this was more, less, or as much as usual; drug consumption (the day before and in general) ; their feeling of hunger before and after the task; and task-related questions about difficulty, monotony, task fatigue, inattentive phases, handedness of the shown actor, subjective guessing rate for the answered questions during the task, recognizability of the objects and actions, change in individual segmentation strategy within-session and between-sessions, and their segmentation strategy.

Behavioral Reliability Measures

Intra-individual Retest Reliability of Unit Marking Responses

As the unit marking procedure is a subjective judgment task and, therefore, responses cannot be right or wrong, retest reliability was assessed on the single-subject as well as on the group level to ensure that responses were consistent and meaningful. Details regarding these reliability

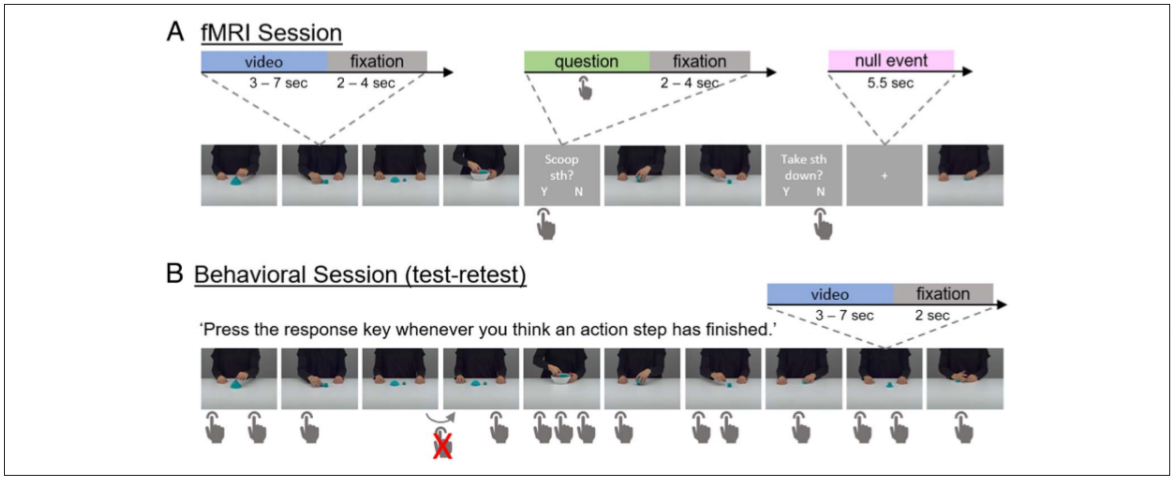

Figure 3. Experimental task design. (A) In the fMRI session, video trials (action video followed by a jittered ISI that showed a white fixation cross) and null event trials (showing a white fixation cross) were passively attended to but question trials (question followed by a jittered ISI that showed a white fixation cross) required participants to confirm or reject an action description with regard to the preceding action video by button press. The question disappeared only after button press and followed 14% of the action videos. For the video trials, here, each single frame represents a full action video plus ISI as indicated by the dotted lines. In total, 308 videos, 42 questions, and 49 null events were presented to each participant, separated in seven blocks with short breaks in between. (B) In the two subsequent behavioral sessions (test–retest), each participant saw the same videos in the same sequence as during fMRI and indicated by button press (hand icon) when they thought an action step had finished. In case no response was given (hand icon crossed out in red), the video at hand was repeated. Participants were instructed to use this mechanism in case they wanted to rewatch the video before indicating action steps. Thus, minimally one button press was necessary per action video but no instruction was given about the expected total number of button presses per action video. Each single frame in the figure represents a full action video plus an ISI that showed a white fixation cross, as indicated by the dotted lines. Example videos are provided in an OSF repository (DOI 10.17605/OSF.IO/MGQSF). The entire stimulus material is available via the Action Video Corpus Münster (AVICOM, https://www.uni-muenster.de/IVV5PSY/AvicomSrv/).

measurements have been previously described (Pomp et al., 2021). As the first step, responses were converted from milliseconds to frames (one frame amounting to a 1000/23 msec segment) to allocate each response to the correspondingly presented frame of the video. Note that we did not subtract any motor RT as participants were highly familiar with the kind of simple everyday actions that we employed, which they saw for the second and third time in the behavioral sessions. Hence, we adopted the premise that responses were delivered in clear anticipation of critical events in the videos, not in a reactive manner.

On the single-subject level, we examined whether test session responses matched retest session responses consistently. To this end, trials with an equal number of responses in the test and retest session were selectively used to define an individual temporal consistency criterion c_i, which was then applied to all trials independent of the number of responses. For each response in each of these equal-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 sessions. In summary, for each retest response t', it was determined whether a test response t appeared within the individual time window around the retest response ($t' \pm c_i$). If this was the case, it was considered a consistent unit marking response. That is, the participant pressed the response key at the same time during the action video in the test and retest session. Subsequently, as a measure of intra-individual retest reliability, the percentage of consistent responses per participant was identified. These consistency rates were statistically compared with the corresponding object study's values using independent-samples t tests and the corresponding Bayesian test with JASP (JASP Team, 2024), and JZS Bayes factors are reported (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

To ensure the validity of our intra-individual retest reliability results, we compared the intra-individual retest reliability results to random button presses. To this end, we extracted the time intervals between button presses (for the first button press in a video, we used the distance to the start of the video) of the test session per participant. From this distribution, we randomly drew and cumulated intervals to simulate random test session data while preserving the stochastic characteristics of the individual behavior. By this procedure, we generated 10 simulated test session data sets, calculated the percentage of consistent responses per participant based on the real retest session data (applying the identical protocol as for the actual behavioral data), and averaged this percentage per participant over the 10 simulations. To test whether the participants performed more reliably than randomly, we calculated a paired-samples t test between the actual percentage of consistent responses per participant and the percentages based on the simulated data sets.

Retest Reliability of Unit Marking Responses at the Group Level

To examine the unit marking responses on the group level, we smoothed the frame-by-frame data of all participants with a rectangular kernel of a width of three frames (3 * (1000/23) ≈ 130.4 msec, 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. Maximally, one response per participant was taken into one bin of three frames so that the total number of participants was the maximum value a bin could reach. 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 r). The r values per video were then Fisher z-transformed, averaged, and retransformed to r to give a mean correlation indicating group-level retest reliability. Furthermore, the r values per video were statistically compared with the corresponding object study's values using independent-samples t tests and the corresponding Bayesian test reporting JZS Bayes factors (Rouder et al., 2009) with JASP.

Group-consistent Unit Mark (M) Determination and Their Relation to TU Events

Determination of Group-consistent Unit Marks

The maximum frame value per video was taken to indicate a group-consistent unit mark (M) as it reflects the point of maximum group agreement. To assure the meaningfulness of these values, we utilized the 10 simulated test session data sets that were generated to evaluate intra-individual retest reliability. We applied the same protocol to these 10 simulated data sets as we did to the original data. Thereby, we determined simulated group-consistent unit marks and then compared their maximum frame values to the actual one per video.

To determine the non-unit-mark (nM) as relevant points in time 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.

Temporal Convergence of Participant-determined Unit Marks and Objective Events

We investigated the temporal relation of Ms to TUs by evaluating whether the majority of Ms coincides with TUs. We examined how often an M was not further than two frames (i.e., maximally ~130 msec) away from a TU. Subsequently, we compared this result to randomly distributed unit marks to validate the systematics of the relationship. Equal to the protocol for the test–retest performance of individual participants, we shuffled the time intervals generated by the unit marks, randomly drew from this shuffled distribution, and cumulated intervals to simulate random unit marks while preserving the stochastic characteristics of the group behavior. This way, we generated 10 simulated unit mark data sets, examined per data set the proportion of simulated Ms being not further than two frames away from a TU, and then calculated a one-sample t test to compare simulated and actual coincidence rates.

Effects of Object–Action Associations on the Temporal Relation of Ms to TUs

Building on our previous study that showed that Ms were systematically delivered in relation to TU events (Pomp et al., 2021), we hypothesized weak object–action associations to increase the temporal proximity of M to TU. As the first analytic step, we tested whether the M-TU difference distributions differed between studies using the Mann–Whitney U test, and tested for equality of variances using Levene's test. Following our hypothesis that Ms are closer to TU events in dough actions (independent of whether they appear before or after the TU), absolute difference values were further analyzed. As these absolute temporal differences between Ms and its closest TUs in both studies had a negative binomial distribution, we fitted a generalized linear (negative binomial) model using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in the R programming language (https://www.R-project.org/). In the model, the absolute temporal differences between Ms and TUs, measured in frames, were predicted by *Study* (i.e., Dough vs. Object) and *Event Type* (i.e., Touch vs. Untouch). In the model, the action categories of the videos were used as a random intercept:

$$absolute(M\text{-}TU) \sim Study \times EventType + (1|ActionCategory). \quad (3)$$

fMRI Data Acquisition

Structural and fMRI data were acquired using a 3-Tesla Siemens Magnetom Prisma MR tomograph with a 20-channel head coil at the Translational Research Imaging Center of the University Hospital Münster. High-resolution, T1-weighted images were obtained by a 3-D-multiplanar rapidly acquired gradient-echo sequence (scanning parameters: 192 slices, repetition time = 2130 msec, echo time = 2.28 msec, slice thickness = 1 mm, field of view = 256 × 256 mm^2, flip angle = 8°). For the functional images, a BOLD contrast was measured by gradient-EPI. Seven EPI sequences were used to measure the seven experimental blocks (scanning parameters: 33 slices, TR = 2000 msec, echo time = 30 msec, slice thickness = 3 mm, field of view = 192 × 192 mm^2, flip angle = 90°).

fMRI Data Analysis

Preprocessing

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

fMRI Design Specification and Whole-brain Statistics

The statistical analyses of the functional images were done using SPM12, implementing a general linear model for serially autocorrelated observations (Worsley & Friston, 1995; Friston et al., 1994) and a convolution with the canonical hemodynamic response function. As regressors of no interest, the six subject-specific rigid-body transformations obtained from realignment were included. 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 subjective action boundaries, as well as objective T and U events, a general linear model was constructed including eight regressors of interest coding for onsets and durations of the specific event types: video trial, group-consistent unit mark of the test–retest session (M), no unit mark in the test–retest session (nM), objective touching event (T), objective untouching event (U), no touching or untouching event (nTU), null event, and question trial. For each of the 340 Ms, an nM was determined (n = 340; see Determination of Group-consistent Unit Marks section) and included in the design. Likewise, all 735 touching and all 808 untouching events were included and correspondingly 735 nTUs (see Video Segmentation and SEC Determination section). Both types of noncritical events (nTU and nM) appeared distributed over the video duration and were chosen to be maximally far away from their corresponding events (TU and M, respectively). The rapid succession of Ms and TUs with naturally jittered interevent intervals made it possible to differentiate associated BOLD responses, and the difference in frequency of occurrence ensured the overall low overlap between M and TU events. Moreover, we applied the post hoc variance inflation factor (VIF) method using the CANlab imaging analysis tools (https://canlab.github.io/) to rule out multicollinearity issues and this yielded VIFs below 10 (object study VIFs < 7.2, dough study VIFs < 7.9), speaking against a severe issue of collinearity.

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 group-determined event boundaries. Analogously, t-contrasts for T versus nTU, U versus nTU, and the complete video trials versus null events were conducted on the first level and then passed to a second-level t test. To elucidate the central question of the object–action association effect, we contrasted activity patterns for play dough actions of this study to activity patterns for object actions of our previous study in a second-level two-sample t test. We did this for all full-length videos as well as time-point specifically at M, T, and U events. Importantly, because we only considered interactions, all contrasts controlled for the main effects of group, action type, and so forth.

To identify brain areas where neural activity was significantly explained by both object and play dough actions' events, we performed conjunction analyses testing against the conjunction null hypothesis, p(false discovery rate [FDR]) < .005 (Nichols, Brett, Andersson, Wager, & Poline, 2005) using a second-level, one-way ANOVA on individual statistical maps derived from the M > nM, T > nTU, U > nTU, and video > null contrasts.

For the second-level, whole-brain analyses, we applied FDR correction at p < .005 peak level and a cluster extent threshold of 15 voxels. Activity patterns were visualized using *bspmview* (DOI 10.5281/zenodo.595175) in MATLAB R2022a, and graphs for visualization were generated using the *ggplot2* library (Wickham, 2016) in RStudio (R Core Team, 2022). We uploaded the unthresholded statistical maps to NeuroVault.org (Gorgolewski et al., 2015), which are available at https://neurovault.org/collections/16065/.

ROI Analyses

To inspect the effects of object–action associations more specifically in the hypothesized regions, we additionally performed planned ROI analyses. Addressing aIPL, we used area PFt (Caspers et al., 2006, 2008) of the Julich-Brain Cytoarchitectonic Atlas (Amunts, Mohlberg, Bludau, & Zilles, 2020; Eickhoff et al., 2005), noting that also relevant peak MNI coordinates of our previous study all fell into this field (Schubotz et al., 2014). The aIPL Julich-Brain ROI was created using the SPM anatomy toolbox (www.fz-juelich.de/inm/inm-7/JuelichAnatomyToolbox). As second ROI, we used the PHC. We defined the extend of the PHC ROI using the Harvard-Oxford anatomical atlas (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases) and the software MRIcron (https://www.mccauslandcenter.sc.edu/mricro/mricron), including voxels if the atlas labeled them as "Parahippocampal Gyrus, posterior division" or "Parahippocampal Gyrus, anterior division" with a probability of >25% (Li et al., 2016; Ward, Chun, & Kuhl, 2013). As third ROI, we employed the temporo-occipital area sensitive to biological motion (BMA), which we gratefully adopted from a recent meta-analysis on the functional

organization of the posterior lateral temporal cortex (Hodgson, Lambon Ralph, & Jackson, 2023). We extracted the mean contrast estimates of our main contrasts for each ROI using the Marsbar toolbox (Brett, Anton, Valabregue, & Poline, 2002), which were then compared between studies by a two-sample t test (unequal variances, $\alpha = .05$, two-sided) per region using MATLAB R2022a.

RESULTS

Behavioral Reliability Measures

Intra-individual Retest Reliability of Unit Marking Responses

Concerning single-subject retest reliability, on average, 63.27% were consistent responses (i.e., the test response matched the retest response in time) ranging between the participants from minimally 48.18% to maximally 71.22% ($SD = 6.34$). The individual consistency criterion c_i, which defined the width of the time window around the retest response separately for each participant, was minimum 3.9 frames (i.e., ~170 msec), median 5.7 frames (i.e., ~248 msec), and maximum 11.3 frames (i.e., ~491 msec). Importantly, the consistency of the participants' unit marking behavior was significantly higher than the consistency of simulated random button presses, $t(31) = 17.81$, 95% CI [28.65, 36.07], $p < .001$, $d = 3.15$, two-sided. Thus, participants' unit marking behavior followed a specific nonrandom pattern and was intra-individually consistent across the test–retest sessions. Compared with the object manipulation study (Pomp et al., 2021), the intra-individual retest reliability was similar regarding the individual percentages of consistent responses as indicated by a Bayesian independent-samples t test that showed evidence for the null hypothesis and its classical counterpart yielding nonsignificant results, $BF_{01} = 3.502$), $t(61) = 0.139$, $p = .89$, $d = 0.035$, two-sided. With regard to the respective comparison to random button presses, a greater Cohen's d of 3.15 in dough study's individual retest reliability versus 1.91 in the object study, indicated that individual participants' segmentations were even more systematic for dough videos.

Retest Reliability of Unit Marking Responses at the Group Level

Corresponding to the single-subject retest reliability results, between-subject unit marking behavior was consistent, as revealed by a highly significant correlation between group-based test–retest segmentation performance. That is, correlations testing the group level retest reliability gave a mean correlation of test and retest smoothed time series of frame values per video of $r_z(292) = .72$ ($r_{min} = .40$, $r_{max} = .90$; each individual correlation per video being significant, all $p \leq .0001$). Compared with the object manipulation study (Pomp et al., 2021), group-level retest reliability was significantly higher for dough manipulations, $t(586) = 17.153$, $p < .001$, $d = 1.415$, two-sided ($BF_{10} = 1.365 \times 10^{+50}$).

Group-consistent Unit Mark (M) Determination and Their Relation to TU Events

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 9.93 ($SD = 2.00$), ranging from 6 to 18. All maximum frame values were at least 2 *SD*s above the mean frame value of the respective video, following previous approaches (Pomp et al., 2021; Schubotz et al., 2012). In contrast, the maximum frame values resulting from simulated random unit markings ranged, on average, between 6.11 and 6.37 (i.e., <9.93). In none of these simulated data sets all maximum frame values passed the criterion of being at least 2 *SD*s above the respective video mean frame value. Taken together, this finding suggests that the participants did not segment the action videos randomly, and overall, the group showed a specific nonrandom segmentation behavior.

Furthermore, we inspected the relation between the number of Ms and the number of TUs per video: The number of Ms per video on group level ranged from one to four ($M = 1.2$, $SD = 0.36$, $n = 294$) and was significantly lower than the number of TUs per video that ranged from three to six ($M = 5.2$, $SD = 1.01$, $n = 294$; $t(586) = 64.97$, 95% CI [3.97, 4.22], $p < .001$, $d = 5.36$, two-sided). On the single-subject level, the average number of individual test–retest consistent unit marking responses per video ranged from 0.6 to 1.9 with a mean of 1.4 ($SD = 0.26$, $n = 294$). Crucially, the number of individually consistent unit marking responses per action significantly correlated with the number of TUs per action video, $r(292) = .55$, $p < .0001$, as well as the number of group-level Ms that positively correlated with the number of TUs, $r(292) = .17$, $p = .003$, both pointing to a systematic relationship between the number of Ms and TUs.

Temporal Convergence of Participant-determined Unit Marks and Objective Events

With regard to the temporal relation of Ms to TUs, for more than one third (39.1%) of the Ms, the time lag to the next TU was maximally two frames, that is, ±130 msec. This coincidence rate was significantly higher than the coincidence rates obtained from the 10 sets of simulated random unit marks, $t(9) = -9.46$, 95% CI [24.32, 30.03], $p < .0001$, $d = 2.99$, two-sided underpinning our expectation that Ms were systematically delivered in relation to TUs. Compared with the object manipulation study (Pomp et al., 2021), this significant coincidence rate's difference to simulated random unit marks was more pronounced in the dough study with a Cohen's d of 2.99 compared with a

Cohen's d of 1.27 in the object manipulation study, indicating a stronger systematicity on the group level.

Effects of Object–Action Associations on the Temporal Relation of Ms to TUs

As shown above and in Pomp and colleagues (2021), single subject as well as group behavior was consistent across test and retest sessions in both studies, and descriptive behavioral values regarding the number of Ms per video were comparable between both studies. Still, and as hypothesized, our current results showed a higher coincidence rate between Ms and TUs, with 39.12% for play dough actions compared with 28.3% for object actions. Furthermore, inspecting the temporal distances between Ms and their closest TUs, Levene's test for equality of variances indicated unequal variances, $F(1, 688) = 5.71, p = .017$, with dough action M-TU distances having a significantly lower variance ($Var = 23.34$) than object M-TU distances ($Var = 37.12$) and thus, as hypothesized, a smaller spread of data. The distributions of M-TU differences differed significantly between studies ($W = 48885.50, p < .001, r = -.18, n = 690$). To test our hypothesis that Ms are temporally closer to TUs when only weak object–action association is present, we compared the absolute temporal delay between the occurrence of M and TU for object and play dough actions. Generalized linear (i.e., negative binomial) modeling showed that although the two studies were not significantly different, Wald $\chi^2(1) = 0.02, z$ test $= -0.02$, $p = .89, d = 0.02$; dough: mean = 4.1 ± 3.2, median = 3.5; object: mean = 5.5 ± 4.4, median = 5, generally, the M-T differences differed significantly from the M-U differences, Wald $\chi^2(1) = 13.87, z$ test $= 0.30, p < .001, d = 0.30$; M-T: mean = 4.84 ± 4.0, median = 4; m-u: mean = 4.78 ± 3.7, median = 4. Furthermore, a significant interaction between *Event Type* (touch, untouch) and *Study* (dough, object) was observed, Wald $\chi^2(1) = 15.98, z$ test $= -0.48$, $p < .001, d = -0.48$. To elucidate this interaction, we conducted Bonferroni-adjusted post hoc contrasts, which revealed that although the M-T differences were significantly different between the two studies, z-ratio(object/dough) $= -3.41, p < .001$ (Figure 4B), the M-U differences were not significantly different, z-ratio(object/dough) $= 0.13, p = .89$ (Figure 4D). These results indicate that actions were segmented closer to T events in case of weak object–action associations. For signed and unsigned M-T and M-U differences, see Figure 4. The signed temporal differences in Figure 4A and Figure 4C illustrate when participant-judged Ms appear in relation to T and U events. Moreover, the unsigned differences shown in Figure 4B and Figure 4D address the question whether Ms were temporally closer to T or U events independent of the sign.

fMRI Results

To investigate the whole-brain and ROI effect of object–action associations, we compared brain activity patterns of the two studies for the full video length (video > null)

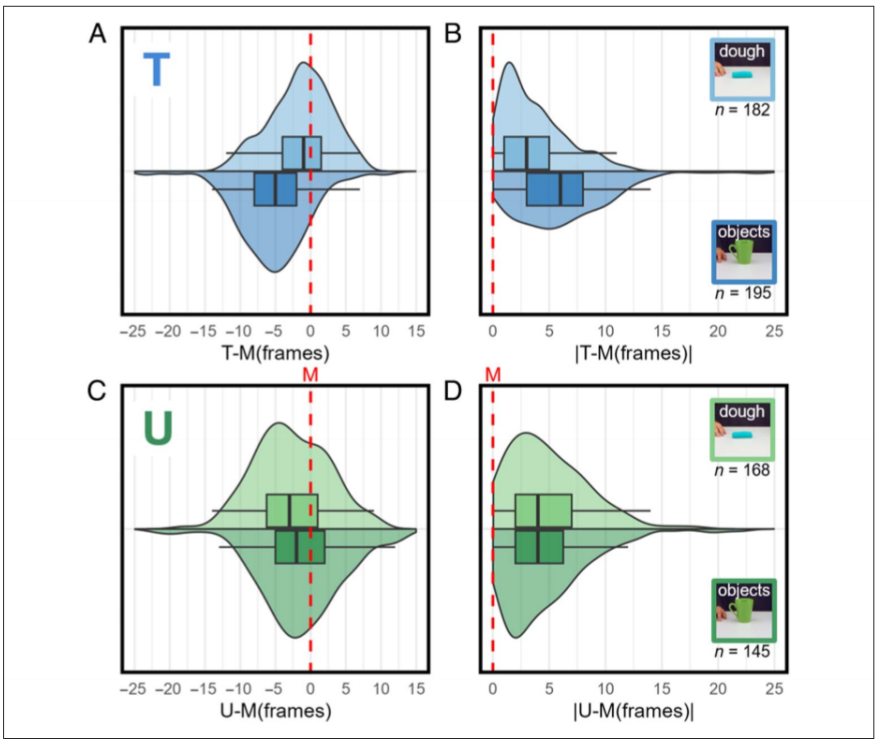

Figure 4. The temporal relation of unit marks (M) to touch (T) and untouch (U) events. The distribution of M to T differences (blue) shown as signed values (A) and unsigned values (B) given in frames and grouped by study (top, light: dough study; bottom, dark: object study). Similarly, the distribution of M to U differences (green) shown as signed values (C) and unsigned values (D) grouped by study. The red dashed line at $x = 0$ indicates when participants behaviorally segmented actions, that is, a unit mark (M) was determined. The action videos had a frame rate of 23 frames per second (1 f ≙ 43.5 msec).

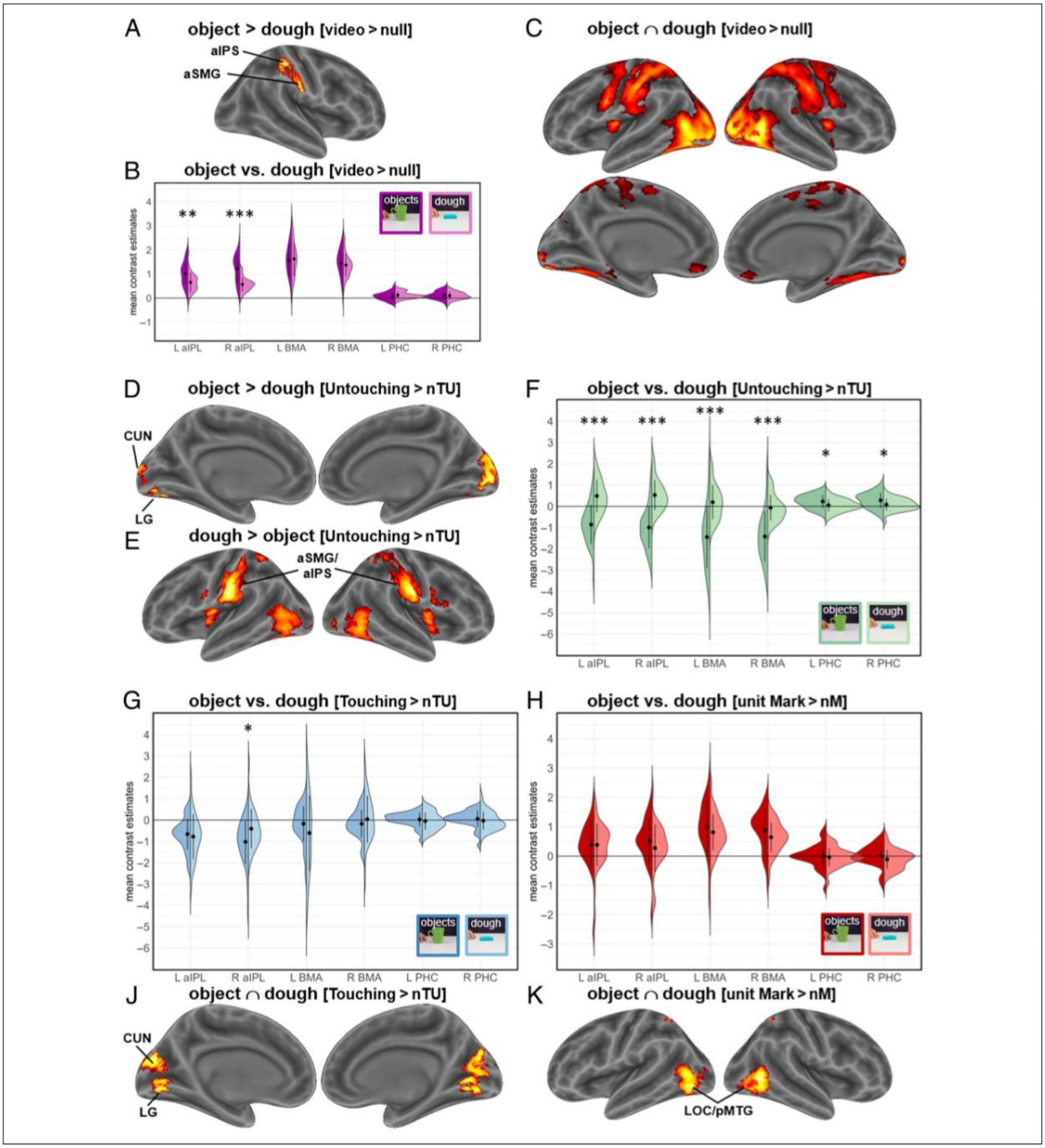

Figure 5. fMRI activation in contrasts and conjunctions between object and dough data at $p < .005$, peak-level FDR-corrected, and ROI analyses of left (L) and right (R) aIPL, biological motion area (BMA), and PHC. A, B, and C illustrate the between-studies' effects for the full video length (video > null; purple). D and E show the whole-brain effects, and F shows the ROI analyses for the between-study comparison at untouching events (U > nTU; green). ROI analyses for the between-study comparison at touching events (T > nTU; blue) are illustrated in G and for unit marks (M > nM; red) in H. Finally, between-study conjunction results are depicted in J for touching events and K for unit marks. For ROI analyses: Mean contrast estimates were extracted from the contrasts video > null, U > nTU, T > nTU, and M > nM of the object (dark shade) and dough (light shade) study. Note that all comparisons show Group × Event interaction effects. For objects $n = 31$, for dough $n = 33$. Statistics: two-sample t tests (*two-tailed*). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Unthresholded statistical maps of the whole-brain analyses have been uploaded to NeuroVault.org and are available at https://neurovault.org/collections/16065/.

Table 1. Maxima of Activation from the Contrasts and Conjunctions of Dough Study and Object Study Contrasts at $p < .005$ Peak-level FDR-Corrected

					MNI Coordinates		
Macroanatomical Location	*Abbreviation*	*H*	*Cluster Extent*	*t Value*	*x*	*y*	*z*
U > nTU							
Dough > object							
Anterior supramarginal gyrus/ ventral postcentral sulcus	aSMG/vPoS	R	672	9.48	60	−16	26
Anterior supramarginal gyrus	aSMG	R		9.38	66	−16	29
Anterior intraparietal sulcus	aIPS	R		7.19	57	−25	53
Anterior supramarginal gyrus/ ventral postcentral sulcus	aSMG/vPoS	L	742	9.17	−60	−19	23
Anterior intraparietal sulcus	aIPS	L		7.27	−57	−25	50
Superior parietal lobule	SPL	L		5.82	−18	−49	71
Mid-insula	MIC	L	152	7.94	−39	−4	14
		R	162	7.08	39	−1	14
Insula	IC	R		6.48	39	−1	−1
Lateral occipito-temporal cortex	LOTC	R	341	7.17	51	−70	−7
Posterior inferior temporal gyrus	pITG	R		6.36	51	−58	−19
Lateral occipito-temporal cortex	LOTC	L	379	7.16	−48	−73	−1
Ventral precentral gyrus	preCG	R	127	5.70	57	11	35
		L	23	4.30	−57	8	29
Cerebellum	CER	L	25	5.39	−15	−67	−46
Object > dough							
Lingual gyrus	LG	R	445	7.18	15	−88	2
		L		6.29	−24	−76	−4
Cuneus	Cun	R		6.52	9	−94	11
		L		6.22	−9	−100	14
Object ∩ dough							
Parahippocampal cortex	PHC	R	18	6.37	33	−55	−7
		L	7	5.37	−33	−55	−7
Dorsal premotor cortex	PMd	L	29	5.48	−21	−10	56
Video > null							
Object > dough							
Anterior intraparietal sulcus/ postcentral sulcus	aIPS/PoS	R	174	6.49	42	−31	47
Anterior supramarginal gyrus	aSMG	R		6.00	60	−19	32
Object ∩ dough							
Posterior middle temporal gyrus	pMTG	R	1834	15.98	48	−64	2
Inferior occipital gyrus	IOG	R		15.09	42	−73	−7
Middle occipital gyrus	MOG	R		14.30	30	−91	5

Table 1. (*continued*)

					MNI Coordinates		
Macroanatomical Location	*Abbreviation*	*H*	*Cluster Extent*	*t Value*	*x*	*y*	*z*
Hippocampus	HC	R		5.01	24	−13	−16
Posterior middle temporal gyrus	pMTG	L	1665	15.24	−45	−67	5
Lingual gyrus	LG	L		13.12	−27	−91	−10
Inferior occipital gyrus	IOG	L		12.60	−39	−76	−7
Fusiform gyrus	FG	L		11.20	−39	−61	−13
Parahippocampal gyrus	PHG	L		3.90	−24	−28	−16
Insula	IC	L	4379	10.79	−36	−7	14
Ventral postcentral sulcus	vPoS	L		10.79	−51	−25	41
Ventral premotor cortex	PMv	L		10.70	−57	5	32
Insula	IC	R	4379	10.37	36	−4	14
Postcentral gyrus	PoG	R		10.16	54	−19	41
Anterior intraparietal sulcus	aIPS	L		9.94	−42	−31	47
Cerebellum	CER	L	117	7.97	−9	−73	−43
		R	89	6.44	12	−73	−43
Rectal gyrus	RG	L	105	6.02	0	29	−22
Mid cingulum	MCC	R	22	4.72	15	−16	44
SMA	SMA	L	81	4.63	−9	−1	56
		R		4.44	9	2	56
Amygdala	AMY	R	29	4.34	36	−1	−16
M > *nM*							
Object ∩ dough							
Lateral occipital cortex	LOC	L	252	8.41	−48	−73	−7
		L		8.02	−45	−70	2
Posterior middle temporal gyrus	pMTG	R	274	8.29	48	−64	2
Superior parietal lobule	SPL	R	40	5.55	18	−58	68
		L	49	5.24	−21	−58	65
T > *nTU*							
Object ∩ dough							
Cuneus	CUN	L	657	6.75	−6	−82	23
		R		5.99	9	−79	26
Lingual gyrus	LG	L		6.17	−9	−76	−1
		R		5.87	12	−76	−4

H = hemisphere; L = left; R = right; U = untouching events; nTU = non-(un-)touching events; M = unit marks; T = touching events.

as well as for the time-point-specific activation contrasts at M (M > nM), T (T > nTU), and U (U > nTU) events. Please note that we always refer to the just enumerated contrasts when we refer to M, T, and U as events. This means that all reported between-study time-point-specific effects are interaction effects (e.g., object study M > nM vs. dough study M > nM), ruling out group effects.

The Entire-video Effects

Comparing object versus dough videos for the full video length (Figure 5A), we found a single cluster in the right aIPL to be significant, including the posterior bank of the ventral postcentral sulcus, the anterior supramarginal gyrus (aSMG), and the anterior intraparietal sulcus (aIPS). ROI analyses affirmed and extended this result by yielding significant activation increases not only in the right, $t(44.53) = 5.77$, $p < .001$, $d = 1.45$, *two-tailed*, but also in the left, $t(58.62) = 3.30$, $p = .002$, $d = 0.82$, *two-tailed*, aIPL for strong object–action associations in object manipulations (Figure 5B). The reverse contrast did not yield significant results. For the common activity between studies during the entire action videos, see the corresponding conjunction results as illustrated in Figure 5C and Table 1.

Interaction Effects at Specific Time Points in the Video (M, T, U)

Contrasting object versus dough videos at critical time points' contrasts, there were no significant differences at M or T events but at U events (Figure 5D) in bilateral cuneus and lingual gyrus. Moreover, ROI analyses (Figure 5F) showed increased activity in bilateral PHC, left PHC: $t(60.96) = 2.43$, $p = .018$, $d = 0.60$, *two-tailed*; right PHC: $t(56.86) = 2.53$, $p = .014$, $d = 0.63$, *two-tailed*.

The opposite contrast of dough versus object videos (Figure 5E) led to higher BOLD responses at U events in bilateral aIPS extending into the SMG in the right hemisphere and to the superior parietal lobule (SPL) in the left hemisphere; furthermore, bilateral insula, bilateral lateral occipital cortex (LOC), and bilateral ventral precentral gyrus activations were detected. The ROI analyses (Figure 5F) showed dough versus object effects at U events in the bilateral aIPS (left: $t(58.75) = 6.35$, $p < .001$, $d = 1.58$, *two-tailed*; right: $t(54.27) = 7.06$, $p < .001$, $d = 1.76$, *two-tailed*) and bilateral BMA, left: $t(45.87) = 5.51$, $p < .001$, $d = 1.39$, *two-tailed*; right: $t(44.01) = 5.68$, $p < .001$, $d = 1.43$, *two-tailed*. Moreover, comparing dough versus object yielded a significant increase in the right aIPS ROI for T events, $t(60.00) = 2.51$, $p = .015$, $d = 0.62$, *two-tailed* (Figure 5G), whereas whole-brain contrasts at T events were nonsignificant. Neither whole-brain nor ROI analyses revealed significant differences between dough and object videos' time-point-specific M activity (for ROI analyses, see Figure 5H).

Conjunction Effects at Specific Time Points in the Video (M, T, U)

To examine whether dough video effects at M, T, and U resemble corresponding object video effects, conjunctions between studies were calculated, and they generally replicated T- and M-specific activity patterns. For T, the conjunction yielded bilateral cuneus as well as bilateral lingual gyrus activity (Figure 5J), and for M, the conjunction revealed bilateral LOC and bilateral SPL activation (Figure 5K). Notably, the U-specific activity was partially replicated. The conjunction showed overlapping activity in bilateral PHC and left lateralized dorsal premotor area. See Table 1 for the peak maxima of the described contrasts and conjunctions.

DISCUSSION

Previous studies have shown that motion information is of central importance for the brain segmentation of observed actions. Accordingly, we recently showed that touching–untouching events indicating maximal motion changes are an efficient cue for participant-judged event boundaries and are associated with specific processing steps at the neural level (Pomp et al., 2021). In the current fMRI study, we hypothesized that objects also have a significant influence on action segmentation because they are associated with specific manipulations. Extending the previous study, we replaced objects with formed pieces of dough to weaken the object–action associations and compared the behavioral and neural processes of action segmentation between the two fMRI studies. Findings show that, indeed, objects influence action segmentation behavior and the neural processing at specific events.

Behavioral findings showed that touching–untouching information was used for action segmentation, no matter whether object-associated action knowledge was strong or weak. Moreover, intra-individual and group retest reliability measures corroborated reliable segmentation behavior for both studies, as tested via a unit marking procedure (Newtson, 1973). In both object and dough videos, participants reported event boundaries systematically in relation to (un-)touchings. However, as expected, the variance in segmentation behavior was significantly smaller when object–action associations were weak. In addition, when compared with random button presses, participants' segmentations were even more systematic for dough videos. Accordingly, the retest reliability on the group level was higher. In summary, this suggests a lower dispersion of data values in the absence of strongly learned object–action associations. Besides, behavioral measures of reliability and consistency, as well as event frequencies and systematicity in dough action segmentation, resembled those in object actions, corroborating the interpretability of subjective event boundaries and their systematic relationship to objective touching and untouching events.

Inspecting the temporal relationship between participant-judged event boundaries and (un-)touchings, we observed that, as hypothesized, the coincidence rate between unit marks and (un-)touchings was higher for dough actions. Furthermore, here again, the specific response pattern's coincidence rate differed from simulated random unit marks' coincidence rate more pronounced when object–action associations were weak. This result indicated higher behavioral systematicity in the absence of strong object–action associations. In addition, actions were segmented temporally closer to touching events when object–action associations were weak, indicating increased reliance on objective touching events. This is in line with our previous findings suggesting especially touching events announcing an untouching event to be important anchor points of behavioral action segmentation (Pomp et al., 2021). Note that the systematic relation of Ms to TU does not imply a generally high overlap of events in time, as there were considerably more TU events (735 touching and 808 untouching events) than participant-judged unit marks (340). Thus, consistent with our first study (Pomp et al., 2021), we also found in the dough manipulation study that participant-judged event boundaries very frequently coincided with TU events, but the majority of TU events did not coincide with a participant-judged event boundary. Future studies need to investigate exactly which TU events are used as anchor points triggering subjective boundary detection.

Taken together, the smaller spread of data and the larger behavioral systematicity in the responses to dough videos showed that the subjective event boundaries relied even more on touching events when strong object–action associations are absent. Thus, before having experience-based knowledge of object-associated actions, the individual presumably relies particularly strongly on objective (un-)touchings. In general, our behavioral findings corroborated that relational changes in the form of touchings and untouchings of objects, hands, and ground represent meaningful anchor points in subjective action segmentation. This finding is critical for creating objective event boundaries that can be used for meaningful action segments. Hard and colleagues (2006) underpinned that goal-based event schemas are not required to detect event structure and concluded that physical changes in the actions subserve event segmentation, measured as bursts of change in movement features. Zacks and colleagues (2009) came to a similar conclusion that movement variables play an important role in action segmentation using a motion tracking system and transcribing movement as a set of 15 variables. Notably, both studies agreed that event structure can be extracted from movement parameters but used complex and costly methods to quantify movement. This is not required in our current approach, which illustrates its practical advantage in this area of research.

Extending the picture arising from the behavioral analyses, fMRI data revealed that object information had significant effects on how the brain processes different types of event boundaries. Importantly, based on interaction contrasts from within-study main effects, our approach controlled for mere perceptual differences arising from the sight of objects or dough pieces. We expected that aIPL and PHC processing might be more relevant for the segmentation of object-directed actions than dough-directed actions, whereas the opposite might be true for an area sensitive to biological motion (BMA). Our findings partly confirmed these hypotheses and also revealed that, among the three types of event boundaries, untouchings were associated with prominent differences between object and dough videos. By contrast, modeling brain data with touching events and participant-judged unit marks replicated the effects that we found for object-directed action segmentation largely (see Appendix). We will, therefore, focus our discussions on untouching events. As shown in our previous study (Pomp et al., 2021), participants reported event boundaries in response to a subset of touching–untouching motifs, that is, the point in time where the observed movement increased significantly from null (touching) to positive change (untouching) and thus became highly informative in respect of the upcoming manipulation. We suggest that object–action associations made the biggest difference at untouching events because participants had to rely much more on movement information when observing dough videos as compared with object videos.

At untouching events, activity increased for dough versus object manipulations in the prespecified ROIs aIPL and BMA, along with bilateral insula and bilateral ventral precentral gyrus activity. Conversely, object versus dough manipulations led to increased bilateral activity in the PHC ROI along with bilateral lingual and cuneal activity. These findings corroborated our hypotheses (a) regarding the increased impact of biological motion for action segmentation in the absence of strong object–action associations and (b) regarding the particular role of long-term mnemonic associations of object and context as reflected by parahippocampal sites for action segmentation in the presence of strong object–action associations.

In light of the fact that (un)touching events provide abstracted dynamic information, the BOLD difference in the BMA at untouchings is a strong indication that participants rely heavily on hand movements to meaningfully process action segments in the absence of strong object–action associations. The employed BMA ROI was functionally defined in a recent meta-analysis (Hodgson et al., 2023) for biological motion. Importantly, the reported effect in our study cannot be because of an increase in motion in the stimuli per se because videos differed only with regard to the target of manipulation, dough, or everyday objects. BMA forms part of the ventrodorsal route for visual input (Binkofski & Buxbaum, 2013), which has been argued to process information aconceptually (Mahon, 2023), that is, without "knowing" what the moving object is. Concerning the analysis of critical

events in studies on action observation, participant-judged event boundaries have been found to activate BMAs (Pomp et al., 2021; Schubotz et al., 2012; Speer et al., 2003). Similarly, dough manipulation data showed BMAs to be active at participant-judged unit marks. This unit-mark-related increase was found for both dough and object-directed manipulations but the untouching-related increase was more prominent for dough-related actions. Therefore, the current approach extends our understanding of motion as playing a key role in event structure perception. Because activity in BMA at untouchings was particularly prominent when objects were weakly informative with regard to associated actions, one may speculate that infants' brains at an age when they do not yet have a mature knowledge of object–action associations can already segment actions into meaningful units based on movement information and may even begin to categorize object manipulation types using this structure (Wörgötter et al., 2013, 2020). A similar principle is used to allow robots to gain some kind of "action understanding." These machines are also, without programming them with additional knowledge, agnostic with respect to the action semantics of objects (Ziaeetabar et al., 2021), and (un-)touching sequences (SECs; Aksoy et al., 2011) can be used by them to recognize actions of humans with whom a robot has to cooperate.

Object manipulations that offered associated action options (and thus assumingly an informed predictive action model) showed the hypothesized increase in PHC activity at untouchings. PHC engagement is reliably seen in tasks where contextual associative information is encoded or retrieved from memory (Li et al., 2016; Aminoff et al., 2013) and is sensitive to the stochastic structure of observed events (Schiffer, Ahlheim, Wurm, & Schubotz, 2012; Turk-Browne, Scholl, Johnson, & Chun, 2010; Amso, Davidson, Johnson, Glover, & Casey, 2005). We take the stronger PHC engagement for object versus dough at untouching to reflect a stronger top–down signal of action prediction, as objects contained more information about possible upcoming actions than pieces of dough. This information about possible upcoming actions possibly provided a restriction on the matching process between the observed and the expected action based on object–action association knowledge. In the absolutely reduced scenery we used in our videos, which consisted only of the table surface, one or two objects, and the actress's upper body up to the shoulders (without head/face), contextual-associative information consisted solely in the combination of the respective object(s) and the manipulation performed on it.

Unexpectedly, aIPL activity did not increase for object versus dough videos, but on the contrary, dominated for dough compared with object videos when we modeled brain activity at untouching events. In our view, this result can only be interpreted if we also consider two other conditions in which the same area was also significantly activated: for object versus dough videos when we modeled the entire video length, and for the conjunction of both, object and dough videos in their full length. Thus, the aIPL was not specifically associated with the processing of only object-related information, and its engagement precisely increased at untouchings when weak object–action associations were available. Notably, in our study, untouching is the phase where updating of the current expectation occurs, as reflected by the engagement of frontal, parahippocampal, and insula regions (Pomp et al., 2021). Note that, although this finding was replicated in the present study (see Appendix), here we focus only on the specific modulations of these responses by the strength of object information. Updating expectations would normally mean that object information is used to select a restricted number of possible manipulations, which can be (or are typically) associated with the presented object. Thus, expectations could be restricted based on this kind of long-term memory, as reflected by the dominance of parahippocampal activity for modeling the BOLD response at untouching events for object versus dough videos. However, in the case of dough videos, this restriction was not provided by the piece of dough, and aIPL activity increase must be related to this unrestricted search for expectable manipulations. The aIPL is generally engaged in tasks highlighting object–hand interactions (Pelgrims et al., 2011; Vingerhoets, 2008). The activated cluster in the inferior parietal lobule that we observed included closely colocalized activation maxima in aSMG and aIPS, which have been assigned distinct but synergetic functions underlying the usage of tools. The aSMG was proposed to integrate semantic and technical information about objects, whereas aIPS rather selects the object-appropriate grasp based on object affordances (Bosch et al., 2023). Moreover, the aSMG may be particularly challenged by unfamiliar tools or conflicting alternative object-directed actions, whereas aIPS modulates this competition by structure-based and skilled use knowledge (Bosch et al., 2023; Buxbaum, 2017; Watson & Buxbaum, 2015). In a previous study, we found that activity in the aIPL varied as a function of the number of actions that participants associated with objects or object sets, even when these actions were not observed (Schubotz et al., 2014). Against this background, we suggest that aIPL was observed time-locked to untouchings when object–action associations were weak because of an unrestricted number of candidate actions in the case of dough manipulation videos, reflecting the matching of the beginning manipulation to the large repertoire of possible manipulations unrestricted by object–action associations. In line with this suggestion, Sacheli, Candidi, Era, and Aglioti (2015) demonstrated that the inhibition of aIPL selectively impaired participants' performance during complementary interactions and suggested aIPL to predictively code other people's actions. In addition, Benedek and colleagues (2018) reported that generating new object uses compared with the generation of known object uses was associated with increased left aIPL activation.

Limitations

Although we made every effort to achieve equal experimental conditions for both experiments' samples, we cannot completely rule out that some behavioral differences at the group level are an artifact. To avoid this limitation, future studies need to randomly assign participants to either group. Importantly, this limitation only concerns the comparison of behavioral data as fMRI analyses consisted only of interaction effects that rule out group effects. However, the overall similarity between the behavior of the two experimental groups concerning segmentation frequencies, intra-individual retest reliability, and group coherence as well as the systematic relationship to TU events gives us confidence in the authenticity of our behavioral results.

Furthermore, objects and dough pieces did not only differ with respect to the associated actions and there might be alternative interpretations of the differences in segmentation behavior. Future research is needed to address the degree of object–action associations in dependence on affordance, functional knowledge, object familiarity, and object complexity. It might be promising to parametrically vary the strength of object–action associations and other dimensions of relevance, and assess their impact on the corresponding segmentation behavior.

Conclusion

Having a life-long experience with manipulable objects provides individuals with a huge repertoire of object–action associations, which is used to efficiently predict object-directed actions. In the present study, modeling brain activity with objective and subjective event boundaries, we showed that object information had, indeed, significant effects on how the brain processes these events. In the absence of strong object–action associations, the increased impact of biological motion processing at objective untouching events, as well as the increased impact of contextual associative information when strong object–action associations were present, confirmed our hypotheses. At the same time, aIPL activity increased for weak object–action associations, presumably because of an unrestricted number of candidate actions. Furthermore, when objects were only weakly informative with regard to associated actions, segmentation behavior became even more systematic and tied to touching events. The present study confirms that objective relational changes in the form of touchings and untouchings of objects, hand, and ground represent meaningful anchor points in subjective action segmentation, rendering them objective marks of meaningful event boundaries. Our findings offer interesting insights into the neural segmentation of object-directed action and the significant influence objects have on the processing of different types of event boundaries because of their association with specific manipulations.

APPENDIX

Figure A1 shows the event-related main effects of touching and untouching events as well as of the participant-judged unit marks for object-directed and dough-directed actions. See Table A1 for the activation peaks of the dough manipulation study and Pomp and colleagues (2021) for the activation peaks of the object manipulation study. It gets obvious that object-directed action activation patterns are largely replicated by dough-directed activation, which means that event processing is mostly not modulated by object–action associations. One striking difference is the activation in aIPS/SMG, which was found for unit marks in object-directed actions and for untouching events in dough-directed actions. Direct whole-brain comparison of the contrasts though yielded no significant differences between action types at unit marks just as the

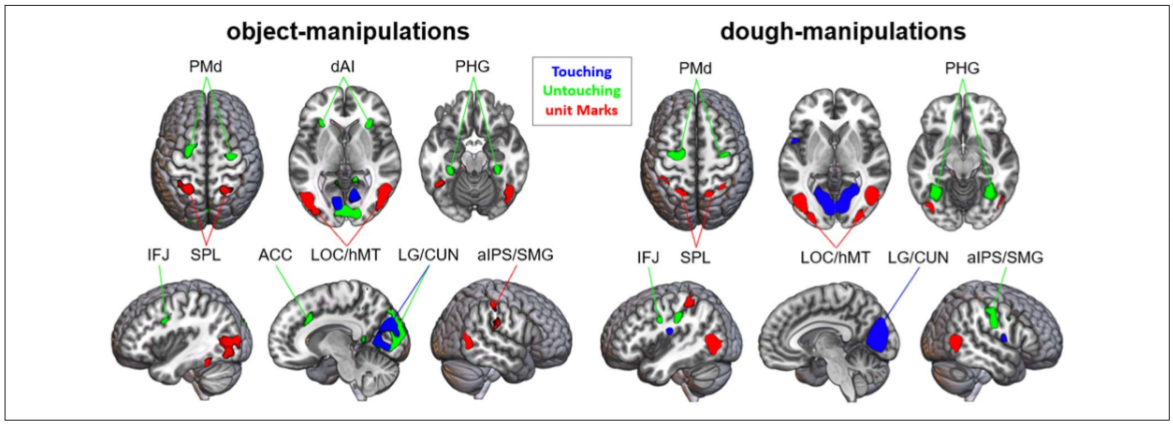

Figure A1. fMRI activation at $p < .005$, peak-level FDR-corrected, for the main contrasts of post-fMRI, participant-judged unit marks (M > nM, red), objective touching events (T > nTU, blue), and objective untouching events (U > nTU, green) of the object-directed action study (left), and the dough-directed action study (right). PMd = dorsal premotor cortex; dAI = dorsal anterior insula; PHG = parahippocampal gyrus; IFJ = inferior frontal junction; LG = lingual gyrus; CUN = cuneus; hMT = motion area; ACC = anterior cingulate cortex.

Table A1. Maxima of Activation from the Main Contrasts of the Second-level Whole-brain Analyses of the Dough Manipulation Study at $p < .005$ Peak-level FDR-Corrected

					MNI Coordinates		
Macroanatomical Location	*Abbreviation*	*H*	*Cluster Extent*	*t Value*	*x*	*y*	*z*
M > nM							
Posterior middle temporal gyrus	pMTG	R	381	10.04	48	−64	2
Inferior occipital gyrus	IOG	R		6.38	30	−97	−1
Lateral occipital cortex	LOC	L	371	9.46	−48	−73	−7
Inferior occipital gyrus	IOG	L		7.01	−27	−97	−1
Superior parietal lobule	SPL	R	101	6.33	33	−52	65
Anterior intraparietal sulcus	aIPS	L	178	6.01	−51	−34	56
Superior parietal lobule	SPL	L		5.74	−24	−58	68
Intraparietal sulcus	IPS	L		5.58	−45	−40	62
Cerebellum	CER	R	19	5.04	9	−73	−43
T > nTU							
Cuneus	CUN	L	1313	9.09	−12	−82	23
Lingual gyrus	LG	L		8.20	−12	−79	8
Calcarine gyrus	CG	L		7.57	−18	−73	14
Lingual gyrus	LG	R		7.47	15	−73	5
Calcarine gyrus	CG	R		7.41	15	−79	17
Cuneus	CUN	R		7.08	15	−79	26
Insula	IC	R	80	7.03	39	−16	23
Rolandic operculum	ROL	L	47	6.25	−42	−16	17
Rolandic operculum (lateral)	ROL	L	33	5.28	−57	5	5
		R	36	5.04	54	−1	8
U > nTU							
Postcentral gyrus/anterior intraparietal sulcus	PoG/aIPS	L	195	7.62	−63	−16	35
Anterior intraparietal sulcus	aIPS	L		4.99	−45	−22	38
		L		4.91	−42	−28	41
Postcentral gyrus	PoG	R	219	7.22	66	−10	29
Anterior intraparietal sulcus	aIPS	R		5.75	51	−16	44
		R		5.05	60	−16	44
		R		4.78	51	−25	44
Mid-insula	mIC	R	62	7.01	36	−4	20
Parahippocampal cortex	PHC	R	114	7.00	36	−58	−7
		L	127	6.66	−36	−55	−10
		L		5.46	−33	−43	−16

Table A1. (*continued*)

					MNI Coordinates		
Macroanatomical Location	*Abbreviation*	*H*	*Cluster Extent*	*t Value*	*x*	*y*	*z*
Cuneus	CUN	R	26	6.44	12	−94	29
Middle intraparietal sulcus	mIPS	L	80	6.13	−27	−43	50
Dorsal premotor cortex	PMd	L	156	6.13	−30	−13	50
Mid-insula	mIC	L	85	6.11	−36	−7	20
Inferior frontal junction	IFJ	L		5.72	−54	2	32
Dorsal premotor cortex	PMd	R	35	5.91	36	−10	56
Middle intraparietal sulcus	mIPS	R	49	5.34	27	−40	50
Posterior intraparietal sulcus	pIPS	L	24	5.02	−21	−73	35

H = hemisphere; L = left; R = right; M = unit mark; nM = non-unit mark; T = touching event; U = untouching event; nTU = non-(un-)touching event.

corresponding ROI analyses (see Results section). At untouching events, however, significant differences were found for aIPS/SMG as well as in other regions as discussed in the Discussion section.

Acknowledgments

The authors thank Monika Mertens and Yuyi Xu for their assistance during data collection. Furthermore, we thank Mina-Lilly Shibata and Simon Reich for their help during the creation of stimulus material, and Ima Trempler for advice regarding data analysis. Finally, we thank Rosari Naveena Selvan for valuable discussions and for proofreading the article.

Corresponding author: Jennifer Pomp, Department of Psychology, University of Münster, Germany, or via e-mail: jennifer.pomp@uni-muenster.de.

Data Availability Statement

The data of the behavioral analyses as well as of the ROI analyses of this article have been deposited in an OSF repository (DOI 10.17605/OSF.IO/MGQSF). Unthresholded statistical maps of all reported and visualized fMRI contrasts in the article have been deposited on NeuroVault (https://neurovault.org/collections/16065/). The entire stimulus material is available via the Action Video Corpus Muenster (AVICOM, 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.

Author Contributions

Jennifer Pomp: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Annika Garlichs: Investigation. Tomas Kulvicius: Formal analysis; Software; Visualization; Writing—Original draft. Minija Tamosiunaite: Conceptualization; Formal analysis; Methodology; Software. Moritz F. Wurm: Methodology; Writing—Review & editing. Anoushiravan Zahedi: Formal analysis; Writing—Review & editing. Florentin Wörgötter: Conceptualization; Funding acquisition; Methodology; Resources; Supervision; Writing—Review & editing. Ricarda I. Schubotz: Conceptualization; Funding acquisition; Methodology; Resources; Supervision; Visualization; Writing—Original draft; Writing—Review & editing.

Funding Information

This research was funded by the German Research Foundation (https://dx.doi.org/10.13039/501100001659; DFG), grant numbers: SCHU 1439/8-1 and WO 388/13-1.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this paper report its proportions of citations by gender category to be: M/M = .373; W/M = .275; M/W = .157; W/W = .196.

REFERENCES

Aguirre, G. K., Mattar, M. G., & Magis-Weinberg, L. (2011). De Bruijn cycles for neural decoding. *Neuroimage*, *56*, 1293–1300. https://doi.org/10.1016/j.neuroimage.2011.02.005, PubMed: 21315160

Aksoy, E. E., Abramov, A., Dörr, J., Ning, K., Dellen, B., & Wörgötter, F. (2011). Learning the semantics of object–action relations by observation. *International Journal of Robotics Research*, *30*, 1229–1249. https://doi.org/10.1177/0278364911410459

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*, 379–390. https://doi.org/10.1016/j.tics.2013.06.009, PubMed: 23850264

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*, 291–298. https://doi.org/10.1016/j.neuroimage.2005.02.035, PubMed: 16061152

Amunts, K., Mohlberg, H., Bludau, S., & Zilles, K. (2020). Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science*, *369*, 988–992. https://doi.org/10.1126/science.abb4588, PubMed: 32732281

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*, 708–717. https://doi.org/10.1111/1467-8624.00310, PubMed: 11405577

Bar, M., Aminoff, E., & Schacter, D. L. (2008). Scenes unseen: The parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. *Journal of Neuroscience*, *28*, 8539–8544. https://doi.org/10.1523/JNEUROSCI.0987-08.2008, PubMed: 18716212

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Benedek, M., Schües, T., Beaty, R. E., Jauk, E., Koschutnig, K., Fink, A., et al. (2018). To create or to recall original ideas: Brain processes associated with the imagination of novel object uses. *Cortex*, *99*, 93–102. https://doi.org/10.1016/j.cortex.2017.10.024, PubMed: 29197665

Binkofski, F., & Buxbaum, L. J. (2013). Two action systems in the human brain. *Brain and Language*, *127*, 222–229. https://doi.org/10.1016/j.bandl.2012.07.007, PubMed: 22889467

Borghi, A. M. (2021). Affordances, context and sociality. *Synthese*, *199*, 12485–12515. https://doi.org/10.1007/s11229-018-02044-1

Bosch, T. J., Fercho, K. A., Hanna, R., Scholl, J. L., Rallis, A., & Baugh, L. A. (2023). Left anterior supramarginal gyrus activity during tool use action observation after extensive tool use training. *Experimental Brain Research*, *241*, 1959–1971. https://doi.org/10.1007/s00221-023-06646-1, PubMed: 37365345

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research*, *206*, 157–165. https://doi.org/10.1016/j.bbr.2009.08.031, PubMed: 19720086

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox [abstract]. In *Paper Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002*. Sendai, Japan.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology*, *76*, 30–77. https://doi.org/10.1016/j.cogpsych.2014.10.001

Buxbaum, L. J. (2017). Distributed neurocognitive mechanisms. *Psychological Review*, *124*, 346–360. https://doi.org/10.1037/rev0000051, PubMed: 28358565

Caspers, S., Eickhoff, S. B., Geyer, S., Scheperjans, F., Mohlberg, H., Zilles, K., et al. (2008). The human inferior parietal lobule in stereotaxic space. *Brain Structure and Function*, *212*, 481–495. https://doi.org/10.1007/s00429-008-0195-z, PubMed: 18651173

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., & Zilles, K. (2006). The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. *Neuroimage*, *33*, 430–448. https://doi.org/10.1016/j.neuroimage.2006.06.054, PubMed: 16949304

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*, 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034, PubMed: 15850749

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. *Journal of Cognitive Neuroscience*, *32*, 326–337. https://doi.org/10.1162/jocn_a_01480, PubMed: 31617822

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*, 429–437. https://doi.org/10.1016/j.neuroimage.2017.11.047, PubMed: 29175612

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. *Human Brain Mapping*, *2*, 189–210. https://doi.org/10.1002/hbm.460020402

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., et al. (2015). NeuroVault.Org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*, 8. https://doi.org/10.3389/fninf.2015.00008, PubMed: 25914639

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*, 586–604. https://doi.org/10.1037/a0024310, PubMed: 21806308

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory and Cognition*, *34*, 1221–1235. https://doi.org/10.3758/BF03193267, PubMed: 17225504

Hodgson, V. J., Lambon Ralph, M. A., & Jackson, R. L. (2023). The cross-domain functional organization of posterior lateral temporal cortex: Insights from ALE meta-analyses of 7 cognitive domains spanning 12,000 participants. *Cerebral Cortex*, *33*, 4990–5006. https://doi.org/10.1093/cercor/bhac394, PubMed: 36269034

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*, e0134316. https://doi.org/10.1371/journal.pone.0134316, PubMed: 26218102

Hunnius, S., & Bekkering, H. (2010). The early development of object knowledge: A study of infants' visual anticipations during action observation. *Developmental Psychology*, *46*, 446–454. https://doi.org/10.1037/a0016543, PubMed: 20210504

JASP Team. (2024). *JASP* (Version 0.18.3). [Computer software]. https://jasp-stats.org/

Kurby, C. A., & Zacks, J. M. (2018). Preserved neural event segmentation in healthy older adults. *Psychology and Aging*, *33*, 232–245. https://doi.org/10.1037/pag0000226, PubMed: 29446971

Li, M., Lu, S., & Zhong, N. (2016). The parahippocampal cortex mediates contextual associative memory: Evidence from an fMRI study. *BioMed Research International*, *2016*, 9860604. https://doi.org/10.1155/2016/9860604, PubMed: 27247946

Mahon, B. Z. (2023). Higher order visual object representations: A functional analysis of their role in perception and action. In G. G. Brown, B. Crosson, K. Y. Haaland, & T. Z. King (Eds.), *APA handbook of neuropsychology: Neuroscience and neuromethods* (pp. 113–138). American Psychological Association. https://doi.org/10.1037/0000308-006

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*, 28–38. https://doi.org/10.1037/h0035584

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*, 436–450. https://doi.org/10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*, 847–862. https://doi.org/10.1037/0022-3514.35.12.847

Newtson, D., Hairfield, J., Bloomingdale, J., & Cutino, S. (1987). The structure of action and interaction. *Social Cognition*, *5*, 191–238. https://doi.org/10.1521/soco.1987.5.3.191

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage*, *25*, 653–660. https://doi.org/10.1016/j.neuroimage.2004.12.005, PubMed: 15808966

O'Neal, C. M., Ahsan, S. A., Dadario, N. B., Fonseka, R. D., Young, I. M., Parker, A., et al. (2021). A connectivity model of the anatomic substrates underlying ideomotor apraxia: A meta-analysis of functional neuroimaging studies. *Clinical Neurology and Neurosurgery*, *207*, 106765. https://doi.org/10.1016/j.clineuro.2021.106765, PubMed: 34237682

Pelgrims, B., Olivier, E., & Andres, M. (2011). Dissociation between manipulation and conceptual knowledge of object use in the supramarginalis gyrus. *Human Brain Mapping*, *32*, 1802–1810. https://doi.org/10.1002/hbm.21149, PubMed: 21140435

Pomp, J., Heins, N., Trempler, I., Kulvicius, T., Tamosiunaite, M., Mecklenbrauck, F., et al. (2021). Touching events predict human action segmentation in brain and behavior. *Neuroimage*, *243*, 118534. https://doi.org/10.1016/j.neuroimage.2021.118534, PubMed: 34469813

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 2022.07.1). [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225, PubMed: 19293088

Sacheli, L. M., Candidi, M., Era, V., & Aglioti, S. M. (2015). Causative role of left aIPS in coding shared goals during human–avatar complementary joint actions. *Nature Communications*, *6*, 7544. https://doi.org/10.1038/ncomms8544, PubMed: 26154706

Sargent, J. Q., Zacks, J. M., & Bailey, H. R. (2015). Perceptual segmentation of natural events: Theory, methods, and applications. In R. R. Hoffman, P. A. Hancock, M. W. Scerbo, R. Parasuraman, & J. L. Szalma (Eds.), *The Cambridge handbook of applied perception research* (Vol. 1, pp. 443–465). Cambridge University Press. https://doi.org/10.1017/CBO9780511973017.029

Schiffer, A. M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (2012). Surprised at all the entropy: Hippocampal, caudate and midbrain contributions to learning from prediction errors. *PLoS One*, *7*, e36445. https://doi.org/10.1371/journal.pone.0036445, PubMed: 22570715

Schubotz, R. I., Korb, F. M., Schiffer, 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*, 1195–1205. https://doi.org/10.1016/j.neuroimage.2012.04.008, PubMed: 22521252

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. *Frontiers in Psychology*, *5*, 636. https://doi.org/10.3389/fpsyg.2014.00636, PubMed: 25009519

Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience*, *3*, 335–345. https://doi.org/10.3758/CABN.3.4.335, PubMed: 15040553

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*, 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010, PubMed: 20720125

Vingerhoets, G. (2008). Knowing about tools: Neural correlates of tool familiarity and experience. *Neuroimage*, *40*, 1380–1391. https://doi.org/10.1016/j.neuroimage.2007.12.058, PubMed: 18280753

Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *Journal of Neuroscience*, *33*, 14749–14757. https://doi.org/10.1523/JNEUROSCI.4889-12.2013, PubMed: 24027275

Watson, C. E., & Buxbaum, L. J. (2015). A distributed network critical for selecting among tool-directed actions. *Cortex*, *65*, 65–82. https://doi.org/10.1016/j.cortex.2015.01.007, PubMed: 25681649

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Version 3.4.0). [Computer software]. New York: Springer-Verlag. https://ggplot2.tidyverse.org

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 Transactions on Autonomous Mental Development*, *5*, 117–134. https://doi.org/10.1109/TAMD.2012.2232291

Wörgötter, F., Ziaeetabar, F., Pfeiffer, S., Kaya, O., Kulvicius, T., & Tamosiunaite, M. (2020). Humans predict action using grammar-like structures. *Scientific Reports*, *10*, 3999. https://doi.org/10.1038/s41598-020-60923-5, PubMed: 32132602

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—Again. *Neuroimage*, *2*, 173–181. https://doi.org/10.1006/nimg.1995.1023, PubMed: 9343600

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*, 651–655. https://doi.org/10.1038/88486, PubMed: 11369948

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*, 201–216. https://doi.org/10.1016/j.cognition.2009.03.007, PubMed: 19497569

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind–brain perspective. *Psychological Bulletin*, *133*, 273–293. https://doi.org/10.1037/0033-2909.133.2.273, PubMed: 17338600

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*, 80–84. https://doi.org/10.1111/j.1467-8721.2007.00480.x, PubMed: 22468032

Zhao, L. (2019). The role of the action context in object affordance. *Psychological Research*, *83*, 227–234. https://doi.org/10.1007/s00426-018-1002-y, PubMed: 29610980

Ziaeetabar, F., Pomp, J., Pfeiffer, S., El-Sourani, N., Schubotz, R. I., Tamosiunaite, M., et al. (2021). Using enriched semantic event chains to model human action prediction based on (minimal) spatial information. *PLoS One*, *15*, e0243829. https://doi.org/10.1371/journal.pone.0243829, PubMed: 33370343

3.3 Study III: Touching-Untouching Patterns Organize Action Representation in the Inferior Parietal Cortex.

Running title:

3.3 TU Patterns Organize Action Representation

Jennifer Pomp, Moritz F. Wurm, Rosari N. Selvan, Florentin Wörgötter,

& Ricarda I. Schubotz (2025)

Neuroimage, 310, 121113

https://doi.org/10.1016/j.neuroimage.2025.121113

Associated online data:

https://doi.org/10.17605/OSF.IO/9V3CQ (OSF repository)

https://www.uni-muenster.de/IVV5PSY/AvicomSrv (Stimulus material on AVICOM)

NeuroImage 310 (2025) 121113

Contents lists available at ScienceDirect

NeuroImage

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

Touching-untouching patterns organize action representation in the inferior parietal cortex

Jennifer Pomp [a,b,*], Moritz F. Wurm [c], Rosari N. Selvan [a,b], Florentin Wörgötter [d], 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] Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Italy
[d] Institute for Physics 3 – Biophysics and Bernstein Center for Computational Neuroscience, (BCCN), University of Göttingen, Germany

ARTICLE INFO

Keywords:
Action observation
Representational Similarity Analysis
Inverse MultiDimensional Scaling
Semantic Event Chain
aIPL
Object-directed action

ABSTRACT

At an abstract temporospatial level, object-directed actions can be described as sequences of touchings and untouchings of objects, hands, and the ground. These sparse action codes can effectively guide automated systems like robots in recognizing and responding to human actions without the need for object identification. The aim of the current study was to investigate whether the neural processing of actions and their behavioral classification relies on the action categorization derived from the touching-untouching structure. Here we show, using a representational similarity analysis of functional MRI data from two experiments, that action representations in left anterior intraparietal sulcus (aIPS) are particularly associated with this categorization of touching-untouching structures. Within the examined action observation network, only the touching-untouching category model selectively correlated with the representational profile of the left aIPS. The behavioral results showed a significant relation between the touching-untouching structure and the observers' judgments on the similarity of actions with weakly-informative objects. Extending prior research on touchings and untouchings as meaningful anchor points for explicit action segmentation, our findings suggest that touching-untouching sequences serve as an organizing principle in inferior parietal action representation.

1. Introduction

Action recognition is crucial in many modern applications like video surveillance, human-computer interaction, web-video search and retrieval, robotics, elderly care, and sports analytics (Herath et al., 2017). Accordingly, automatic action recognition is a popular and promising field of basic and applied research. However, despite having similarities to static image analysis, video data analysis is far more complicated (Jiao et al., 2022). The main challenges of this continuous process arise due to the movement of objects and changes in perspective leading to changing size and appearance, blurriness, and changing light intensities. Moreover, current deep learning networks still depend heavily on extensive pretraining (Han et al., 2021), and object recognition remains particularly difficult due to the complexity and variability within object categories (Liang and Wan, 2020). Therefore, the ultimate goal is to enable machines to learn actions directly from video observations without human intervention. Interestingly, while these tasks pose significant challenges for machines, the human brain excels at action recognition effortlessly, and researchers are actively exploring the underlying mechanisms that enable this ability.

In the attempt to let a robot recognize manipulations performed by a human and let it execute these itself, Aksoy et al. (2011) developed the concept of semantic event chains (SECs). This approach formalizes object-directed actions as sequences of relational changes in the form of touchings (T) and untouchings (U) of objects, hands, and the ground (TUs, hereafter). Thus, SECs or TU sequences encode the contact states between surfaces, without prioritizing the hand over objects or the table surface. Most importantly, robots can recognize and execute manipulations using this SEC-based representation without prior object knowledge (Aksoy et al., 2011). It showed that computer vision using the SEC approach was able to distinguish between 30 different one-handed object manipulations typical of everyday life (Wörgötter et al., 2013). Against this background, the question arose whether these TUs, that are useful for robots to recognize actions, could be also used in

* Correspondence author at University of Münster, Fliednerstrasse 21, Münster 48149, Germany.
E-mail address: jennifer.pomp@uni-muenster.de (J. Pomp).

https://doi.org/10.1016/j.neuroimage.2025.121113
Received 20 November 2024; Received in revised form 31 January 2025; Accepted 3 March 2025
Available online 9 March 2025

human action recognition. And indeed, previous research has shown that TUs are meaningful anchor points for individual action segmentation (Pomp et al., 2024, 2021). Thus, TUs have been shown to be objective and meaningful event boundaries, and certain sequences of TUs emerged as particularly relevant for individuals to determine action steps (Pomp et al., 2024, 2021). Moreover, we found that the observation of Ts and the observation of Us were each associated with different brain activation patterns, emphasizing their importance in the analysis of observed actions.

Interestingly, taking the entire TU sequence of an action into account, also action categories arise from the SEC formalism. For example, turning, pulling, and poking an object all share the same TU sequence and belong to the action category termed "rearrange". Their TU sequence differs from, for example, breaking, ripping-off, and uncovering by picking and placing, that also share a similar TU sequence and build another category termed "break" (Wörgötter et al., 2013). In the current study, we built on these findings and investigated whether the action categories derived from the TU structure of an action might be informative for the brain, and whether they are also reflected in categorization behavior. Specifically, the question was whether there are brain regions that reflect action categories as predicted by their full TU sequence and whether behaviorally determined categories resemble them. To address this question, we employed fMRI-based representational similarity analyses as well as inverse multidimensional scaling (MDS).

Based on the SEC-based ontology of manipulation actions, video stimuli were recorded that belonged to six separate action categories. In two separate experiments, two non-overlapping groups of participants passively watched these videos during the MRI scan. Importantly, the two experiments were similar except for the manipulated objects in the videos. In the first experiment, manipulation actions were performed with daily life objects (e.g., a calculator, a cup, or a piggy bank) and in

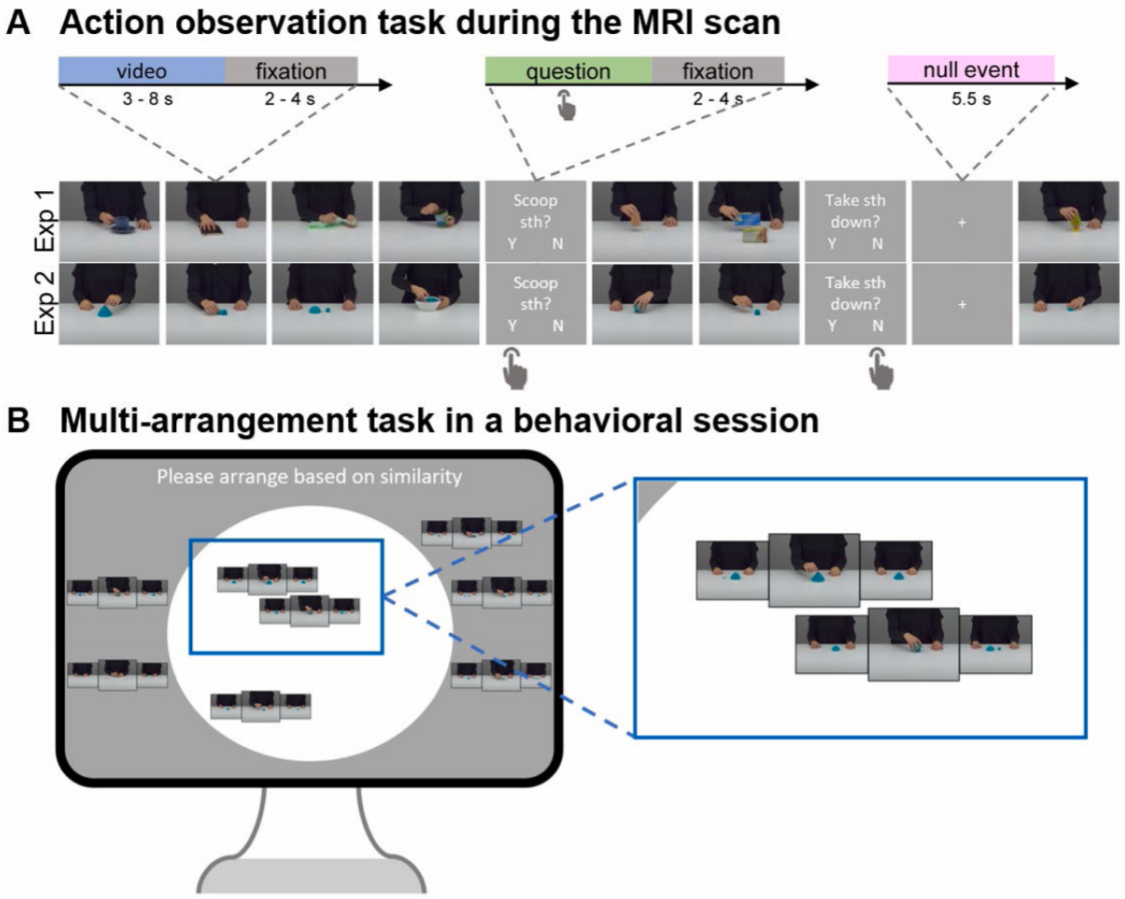

Fig. 1. Experimental task designs. (A) During the fMRI scan, video trials (action videos followed by a jittered inter-stimulus-interval that showed a white fixation cross) and null event trials (showing a white fixation cross) were passively attended to. Question trials (question followed by a jittered inter-stimulus-interval that showed a white fixation cross) required participants to confirm or reject by button press an action description with regard to the preceding action video. The question disappeared only after button press and followed 14% of the action videos. For the video trials, here, each single frame image represents a full action video plus inter-stimulus-interval as indicated by the dashed lines. In sum, 308 videos, 42 questions, and 49 null events were presented to each participant, split into seven blocks with short breaks in between. The task design was equivalent in experiment 1 (Exp 1) and experiment 2 (Exp 2), but the action videos differed, as shown in the lower part of (A). Example videos are provided via the Action Video Corpus Münster (AVICOM, https://www.uni-muenster.de/IVV5PSY/AvicomSrv/), where the entire video stimulus material is available upon request. (B) In a subsequent behavioral session, each participant did a multi-arrangement task with the time course of the action videos being represented by a sequence of three single frame images (see blue box zooming in the items). The participants spatially arranged subsets of these items within a 2D white circular arena on a 27″ computer screen using the computer mouse. (Dis-)similarity between items was expressed through the relative spatial position of the items. The task was implemented in the Meadows web-based platform for psychophysical experiments (http://meadows-research.com). The figure shows an exemplary screen for the evaluation of similarity relationships in the stimulus material of experiment 2.

the second experiment, these objects were replaced by formed play dough pieces that did not resemble meaningful objects (see Fig. 1). This enabled us to focus on the actions' common TU sequences and to control for potential effects associated to object identity. For both experiments, we created a similarity model that captured the action categories as defined by the TU structure given by the SEC framework (referred to as "TU model", hereafter). Furthermore, to investigate to which extent the TU structure of actions influenced how participants subjectively categorize actions, we created a rating-based behavioral similarity model and compared it to the TU model. This model was created using the similarity ratings of a behavioral post-MRI multi-arrangement task that employed the inverse MDS approach by Kriegeskorte and Mur (2012), and we refer to it as "MDS model" hereafter.

With regard to brain activation, we reasoned that if two actions are defined as similar in TU structure, they should evoke similar activation patterns in brain areas coding actions in terms of SEC-like information. Accordingly, we performed a specific form of multivoxel pattern analysis known as representational similarity analysis (RSA) to investigate which brain regions reflect the action categories as predicted by their TU structure. We used a searchlight RSA for a brain-wide analysis and a region of interest (ROI) RSA focused on the action observation network, respectively. Here, we specifically focused on the left anterior inferior parietal lobule (aIPL), as this region has been consistently shown to be sensitive to (observed) hand-object interactions (Murata et al., 2016; Vingerhoets, 2014), observation of touch (Chan and Baker, 2015), physical scene understanding (e.g., how objects rest on each other and how colliding objects behave; Fischer et al., 2016), and reasoning on physical object properties (Reynaud et al., 2016). Moreover, lesion studies have shown that the left parietal lobe plays a major role in apraxia, which comes with severe impairments to associate objects to appropriate actions (Buxbaum and Randerath, 2018). While the aIPL apparently enables a range of closely related functions rather than a single homogeneous one, this profile seemed best suited to represent the TU structure of actions.

2. Methods

For the current investigation, we used fMRI data that was analyzed and published in two previous works (Pomp et al., 2024, 2021). These original datasets included also unpublished data of a post-fMRI multi--arrangement task, which were the focus of the current work. Furthermore, in contrast to the preceding analyses, fMRI-based RSA was used here to identify neural representations of action categories. In several passages in the methods section, we refer to the preceding publications for detailed descriptions. Yet, we repeat details if necessary for immediate understanding of the current study.

2.1. Participants

2.1.1. Experiment 1

As reported in Pomp et al. (2021), 31 participants (M_{age} = 23.84 years, SD = 3.01, age range = 18 - 31 years, 25 women, 6 men) participated in experiment 1. One additional data set was excluded from the analyses as the participant misunderstood the instructions. The participants were all right-handed as determined by the Edinburgh Handedness Inventory (Oldfield, 1971), had intact color perception, normal or corrected-to-normal vision, reported no history of neurological or psychiatric diseases, and self-reportedly met the criteria for MRI scanning. The experiment was conducted according to the Declaration of Helsinki and approved by the local Ethics Committee of the Faculty of Psychology (University of Münster, Germany). The participants provided informed consent and either received course credits (29 of the participants were students of the University of Münster) or were paid for their participation.

2.1.2. Experiment 2

As reported in Pomp et al. (2024), 33 right-handed participants (M_{age} = 23.03 years, SD = 3.06, age range = 18–29 years, 28 women, 5 men) took part in experiment 2. The participants reported having no history of neurological or psychiatric disorders, intact color perception, and had not taken part in related precursor studies. In the course of the behavioral part of the experiment one participant dropped out; hence, this participant's data set was not included in the behavioral model construction but was included in the fMRI data set. The behavioral analysis comprised the data of 32 participants (27 women, 5 men) aged between 18 and 29 years (M_{age} = 22.88, SD = 3.13). The participants gave written informed consent in voluntarily participating in the experiment and were self-reportedly suitable for MRI scanning. They were either paid for their participation or received course credits. The experiment was conducted according to the Declaration of Helsinki and approved by the local Ethics Committee of the Faculty of Psychology (University of Münster, Germany).

2.2. Stimulus material

The object-directed manipulation actions for the video stimuli were chosen according to the SEC framework (Wörgötter et al., 2013). This framework includes transitive actions involving one active hand and one or two objects. Twelve of these actions were selected for the current studies that belonged to six action categories: Rearrange (turn; pull), break (rip off; uncover), destroy (cut; scoop), destruct (take down; take away), construct (put on top; put together), and hide (put over; put into). For experiment 1, each action was recorded using four different objects which resulted in 48 object manipulations (6 action categories x 2 actions x 4 objects). For experiment 2, all 12 actions were performed with formed pieces of blue dough (6 action categories x 2 actions).

Action videos were recorded using an industrial camera and the created video material showed the actress from the front up to the shoulders performing the action on a white table (see Fig. 1). Subsequently, the videos were vertically mirrored so that the actions looked like being performed by the left hand. Each participant saw 50% of the action videos mirrored. For more details on the creation of the video material see Pomp et al. (2021, 2024). In order to control the transition probabilities of the video trials in the experiment, 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 et al. (2011) and NeuroDebian 8.0.0 (Halchenko and Hanke, 2012). See Pomp et al. (2021, 2024) for details.

For the behavioral multi-arrangement task, each action was represented as a sequence of three video images, i.e., single frame images from the start, middle, and end of the action video (Fig. 1B). Therefore, the sequence of images showed the start state, the manipulation, and the end state of the action whereby the middle one was enlarged and highlighted. The image triplets were supposed to represent the course of action of the respective video, that had repeatedly been seen.

2.3. Experimental procedure and tasks

In both experiments, the participants completed three experimental sessions. The first session comprised the MRI session. As described earlier (Pomp et al., 2024, 2021), the participants passively attended to the action videos during the MRI scan. Attention capturing questions followed 14% of the videos asking whether an action description was appropriate for the just seen action video. Participants responded by pressing one of two response keys with their right index and middle finger to reject or affirm the given description. Their response was necessary for the experiment to continue ensuring that participants engaged in attending and recognizing the actions shown in the videos. See Fig. 1A for the experimental trial design. See Pomp et al. (2021, 2024) for details on stimulus presentation in the scanner and the procedure at the scanner.

As second part of the third session, participants did a multi-arrangement task (Fig. 1B). We adapted the multi-arrangement method proposed by Kriegeskorte and Mur (2012), which uses inverse MDS to obtain a distance matrix from multiple spatial arrangements of subsets of items within a 2D space. The participants did this task in a behavioral laboratory using the Meadows web-based platform for psychophysical experiments (http://meadows-research.com). Participants arranged the actions, as represented by a sequence of three single frame images, in a two-dimensional space (on a computer screen of 27″) thereby expressing (dis-)similarity through the relative spatial position of the items. Due to the limited screen size, only a subset of actions was arranged per trial. In experiment 1 the subset included a maximum of 11 actions being simultaneously presented and in experiment 2 a maximum of six actions. These numbers were chosen to give participants enough space to arrange items while having enough items per trial to efficiently gather pairwise similarity ratings in a reasonable total number of trials. In experiment 1, in total the pairwise similarity of 48 actions (12 manipulations x 4 objects) was estimated and in experiment 2, as the object dimension was absent, the pairwise similarity of 12 actions was assessed. Regarding the subset of actions per trial, the concrete items for the second trial's subsets of stimuli (and all subsequent) were determined using an adaptive algorithm that provided the optimal evidence for the pairwise similarity estimates that were inferred from the 2D arrangement of the items on the screen (see Kriegeskorte & Mur, 2012, for details). Therefore, some of the trials included fewer actions than the set maximum, as determined by the algorithm, which allowed participants to refine their judgments within the given arena space. The participants were instructed to drag and drop the stimuli within a circular arena using the computer mouse and place similar actions closer together and dissimilar ones further apart. No explicit instruction was given on which feature to use for similarity. The relative inter-item distances, rather than the absolute screen distances, represented dissimilarities between the items from trial to trial. All items had to be placed in the arena. The task terminated automatically either when a sufficient signal to noise ratio was achieved (the minimum required evidence weight was set to 0.5), or when the maximum session length of 60 minutes was reached. Subsequently, the participants completed a short survey. Please note that the participants knew the action videos very well as they also saw the videos in a behavioral test-retest regime and manually segmented meaningful action steps by button press in the

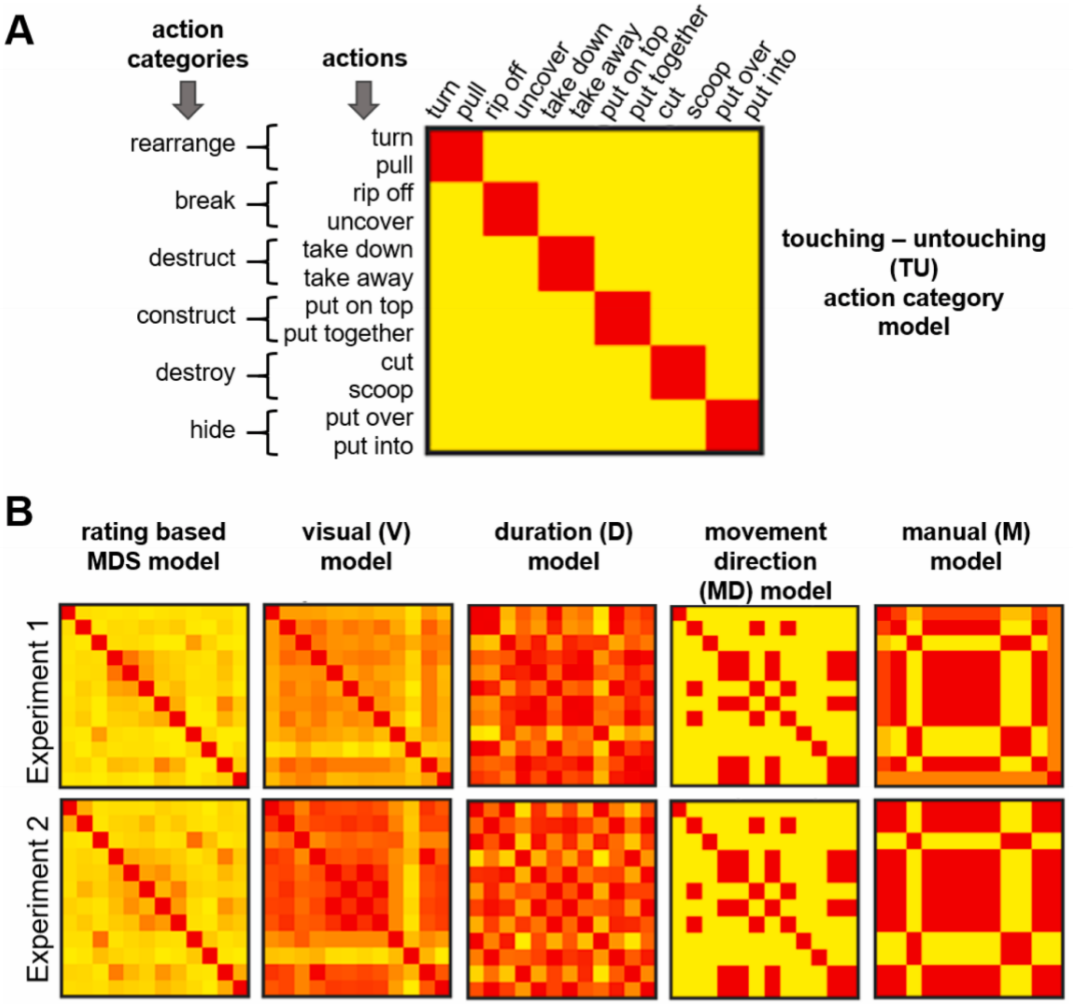

Fig. 2. Model Representational Dissimilarity Matrices (RDMs). (A) The touching-untouching (TU) action category model shown as model RDM with row and column labels, which are representative of the other models. Action categories that contain the concrete actions are given on the left side. (B) Visualizations of the second to sixth model RDMs for experiment 1 (top row) and experiment 2 (bottom row). The scale is from dissimilar (yellow) to similar (red).

second and third experimental session (see Pomp et al., 2021, 2024). We therefore assumed that a triplet representation (i.e., three single frame images) of each action video was sufficient to remind the participants of the respective action video and to let them judge the similarity based on the plot of the action (i.e., the interaction with the object).

2.4. Representational dissimilarity matrices (RDMs)

We created six model RDMs to use them in the RSA (Fig. 2). The first and second models were of main interest for the current study while the third to sixth models were created to control for perceptual stimulus features that might covary with the first and second one. The second model based on participants' judgments. The other models were objective and were derived directly from the video stimuli either through experimenters' judgments (MD model, M model) or feature extraction (V model, D model).

(1) The first model created was a dissimilarity model that captures the action categories as defined by their TU structure according to the SEC framework. We compared each of the 12 actions with each other in terms of whether they belong to the same SEC category (dissimilarity = 0) or not (dissimilarity = 1). This *TU model* was identical for experiment 1 and 2. All following models were created separately for experiment 1 and 2. Importantly, TU categories did not systematically covary with the position of the object relative to the subject. Similarly, the TU categories did not covary with specific grip types, as the same TU sequence was performed with different objects, each requiring a different type of grip.
(2) To investigate to which extent TUs influenced how participants subjectively categorize actions, we created a second model based on similarity ratings. The inverse MDS results from the post-fMRI multi-arrangement task were used for this *MDS model*. The resulting pairwise dissimilarity estimates, i.e., the Euclidean distance between the two actions of a pair, were averaged across trials, resulting in a 48 × 48 (Exp. 1) or a 12 × 12 (Exp. 2) dissimilarity matrix for each participant. Each dissimilarity matrix was normalized by dividing each value by the maximum value of the matrix. Finally, the dissimilarity matrices were averaged across participants and across the exemplars of each of the 12 actions.
(3) To create a low-level visual similarity or *V model*, we computed the pixelwise similarity of each action video. In a first step, we averaged the frames of each action video. We then vectorized the resulting average frames to obtain a pixelwise vector for each action exemplar. The pixelwise action vectors were correlated with each other, resulting in a 12 × 12 correlation matrix. This matrix was transformed into a dissimilarity matrix by subtracting it from 1 (1 – r). Computing pixelwise similarity is a common practice in RSA to capture general low-level visual representations of stimuli (e.g., Kriegeskorte et al., 2008; Peelen and Caramazza, 2012), including videos (de Vries and Wurm, 2023; Wurm and Lingnau, 2015). To capture more specific levels of low-to-midlevel representations upstream of basic pixelwise similarity, several options exist, including FSIM (others are e.g. Radon, silhouette, DNN layers), and specifically for videos, it is also possible to use models based on optical flow and motion energy, as well as more sophisticated models such as Dense Trajectory models and Space-Time-Interest Points (Urgen et al., 2019). However, such higher-level models were beyond the purpose of this study, since we only wanted to control for basic low-level visual similarity of our stimuli.
(4) Since the videos differed in terms of their duration, we computed a video duration or *D model* by computing the absolute difference between video durations.
(5) We created a movement direction or *MD model* that captured whether an object was moved away from the body toward another object or vice versa, i.e., away from another object toward the body. We compared each of the actions with each other with respect to whether the actions were made in the same direction (0 = similar) or in different directions (1 = dissimilar). This model turned out to be similar for experiment 1 and 2. Please note, this model was independent of left-right movements, as these were balanced by showing the stimuli also vertically mirrored (see *Section 2.2. Stimulus material*).
(6) Finally, some of the actions involved both hands, as the non-dominant hand stabilized the object for manipulation, whereas other actions were unimanual, i.e., the non-dominant hand remained still on the table without touching an object. We computed a manual or *M model* by pairwise comparing each of the actions with each other with respect to whether they both were unimanual or bimanual (0 = similar) or one action was unimanual and the other action was bimanual (1 = dissimilar).

Regarding models three to six, each of the perceptual dissimilarity matrices were averaged across the exemplars of each of the 12 actions where applicable.

2.5. Behavioral data analysis

2.5.1. Behavior during the MRI scan

During the MRI scan, attention-maintaining questions irregularly followed the action videos. We evaluated the accuracy and calculated the median reaction time of the correct responses.

2.5.2. Analyses regarding the MDS model

To assess the inter-subject reliability of the MDS models, obtained from the multi-arrangement task, we correlated each participant's pairwise similarities (i.e., the lower triangle of the correlation matrix) with the averaged pairwise similarities of the remaining subjects using leave-one-subject-out cross-validation in MATLAB (https://www.mathworks.com). The resulting correlation values were averaged.

Moreover, to ensure that the MDS models of experiment 1 and 2 reliably captured the behavioral similarity across different types of stimuli (i.e., with and without real objects), we correlated these two models in MATLAB.

Finally, we tested whether the participants' subjective similarity ratings could be explained by the TU model. To this end, we first calculated a multiple regression analysis in JASP (JASP Team, 2024) including the four control models (V, D, MD, and M) as predictor variables, and the subjective similarity ratings as outcome variable. Predictor variables that did not significantly predict the outcome variable were eliminated from the regression equation. Subsequently, the remaining control models were added to the null model to investigate the portion of variance explained by the alternative model which included the TU model as predictor variable.

2.6. fMRI data analysis

2.6.1. fMRI data acquisition, preprocessing, and design specification

For both experiments, MRI data were acquired at the Translational Research Imaging Center (TRIC) of the University Hospital Münster using a 3-Tesla Siemens Magnetom Prisma MR tomograph with a 20-channel head coil. First, high-resolution T1-weighted images were obtained by a 3D-multiplanar rapidly acquired gradient-echo (MPRAGE) sequence (scanning parameters: 192 slices, TR = 2130 ms, TE = 2.28 ms, slice thickness = 1 mm, FoV = 256 × 256 mm2, flip angle = 8°). Subsequently, a blood-oxygen-level-dependent (BOLD) contrast was measured by gradient-echo echoplanar imaging (EPI). Seven EPI sequences measured the seven experimental blocks (scanning parameters: 33 slices, TR = 2000 ms, TE = 30 ms, slice thickness = 3 mm, FoV = 192

× 192 mm2, flip angle = 90°). The fMRI data have been used in two previous papers (Pomp et al., 2024, 2021).

The 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. The preprocessing included slice time correction to the first slice, realignment to the mean image, co-registration of the individual structural scan to the mean functional image, 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 3 mm full-width at half maximum (FWHM) Gaussian kernel. A 128 s temporal high-pass filter was applied to the time-series of functional images to remove low-frequency noise.

The functional images were statistically analyzed using SPM12 applying 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). As regressors of no interest, the six subject-specific rigid-body transformations obtained from realignment were included. To allow for T1-equilibrium effects, the volumes of the first two video presentations of each EPI were discarded. The constructed GLM included 14 regressors of interest coding for onsets and durations of the 12 action video types, null events, and question trials.

2.6.2. Representational similarity analysis (RSA)

The RSA (Kriegeskorte et al., 2008) was carried out using the CoSMoMVPA toolbox (Oosterhof et al., 2016) and the representational similarity toolbox (Nili et al., 2014). For each participant and run, beta weights of the experimental conditions were estimated using design matrices containing predictors of the 12 action conditions, null trials, question trials, and of 6 parameters resulting from 3D motion (translation and rotation) correction. Each predictor was convolved with a dual-gamma hemodynamic impulse response function (Friston et al., 1998). Each trial was modelled as an epoch lasting from video onset to offset. The resulting reference time courses were used to fit the signal time courses of each voxel. The resulting beta weights were averaged across the seven runs to obtain one beta weight per condition and voxel. The searchlight (Kriegeskorte et al., 2006) and ROI RSA were performed in volume space using spherical ROIs with a radius of 12 mm.

2.6.2.1. ROI RSA. For the ROI generalized linear models (GLM) RSA, ROIs were defined in the action observation network based on the peak coordinates of the univariate contrast of all actions vs. null condition (for coordinates see Table 1). While in experiment 2, six bilateral ROIs were defined, experiment 1 did only yield five bilateral and one unilateral ROI (i.e., no univariate peak in right ventral premotor cortex was detected). For each participant, ROI, and condition, we extracted and vectorized the beta values of the ROI to obtain one vector of beta values per action. For each vector, we demeaned the beta values across voxels by subtracting the mean beta value from each individual beta value. Next, we correlated the vectors with each other resulting in a 12 × 12 correlation matrix per ROI and participant. The neural correlation matrices were transformed into a neural RDM (1 – r). The pairwise action comparisons of neural and model RDMs were vectorized, z-scored, and entered as independent and dependent variables, respectively, into a multiple regression RSA. Resulting beta coefficients were entered into a repeated measures analysis of variance (ANOVA) with between-subject factor Experiment(2), and within-subject factors ROI(11), and MODEL(6) to see whether ROIs dissociated before entering the beta coefficients into one-sided signed-rank tests across participants (Nili et al., 2014). Statistical results were corrected for the number of ROIs and tested models (11 ROIs * 6 models = 66 tests) using the false discovery rate (FDR) at q = 0.05 (Benjamini and Yekutieli, 2001).

2.6.2.2. Searchlight RSA. The searchlight GLM RSA was performed using identical parameters as reported above. For all searchlight analyses, individual beta coefficient maps were Fisher transformed and entered into one-sample *t*-tests (Oosterhof et al., 2016). Statistical maps were corrected for multiple comparisons using an initial voxelwise threshold of p = .001 and 10,000 Monte Carlo simulations as implemented in the CoSMoMVPA toolbox (Oosterhof et al., 2016). Resulting z maps were used to threshold statistical maps (at p = .05 at the cluster level), which were projected on a cortex-based aligned group surface for visualization.

To test for multicollinearity between the models, we computed condition indices (CI), variance inflation factors (VIF), and variance decomposition proportions (VDP) using the colldiag function for MATLAB. The results of these tests (Exp. 1: CI<4, VIF<2.2, DVP<0.8; Exp. 2: CI<4, VIF<2.6, DVP<1.0) revealed no indications of multicollinearity that could give rise to potential estimation problems (Belsley et al., 1980).

3. Results

3.1. Behavioral results

3.1.1. Behavior during the MRI scan

To control whether the participants performed the task in the scanner accurately, their performance was analyzed: In experiment 1, participants responded on average in 88.5 % correct (SD = 7.5) with a median reaction time of 1615 ms; and in experiment 2, the mean accuracy was 90.7 % (SD = 6.6) with a median reaction time of 1511 ms.

3.1.2. Results regarding the MDS model

In a first step, we aimed at assessing the reliability of the MDS models obtained from the multi-arrangement task. In both experiments, we observed robust correlations between the ratings of individual subjects: For experiment 1, we found a mean inter-subject correlation of $r(64) = 0.66$, and all but two subjects' data correlated significantly with the group mean (all *ps* < 0.01, *one-sided; for the two remaining subjects:* $p = .09$, and $p = .45$; *one-sided*). For experiment 2, the mean inter-subject correlation was $r(64) = 0.53$, and the data of all the subjects

Table 1
ROI coordinates.

	Experiment 1						Experiment 2					
	L			R			L			R		
ROI	x	y	z	x	y	z	x	y	z	x	y	z
aIPS	-51	-25	41	45	-22	41	-48	-25	41	42	-28	44
LOTC	-45	-70	8	51	-67	-1	-42	-73	-4	48	-64	2
pIPS	-24	-76	35	30	-67	32	-24	-73	32	27	-73	38
SPL	-24	-55	59	30	-49	59	-27	-55	62	21	-64	59
PMd	-27	-10	53	30	-7	59	-24	-7	56	24	-4	53
PMv	-54	5	38	–	–	–	-57	8	29	60	8	32

Note. Spheres of 12 mm. ROI = Region of interest, L = left hemisphere, R = right hemisphere, x, y, z = MNI coordinates, aIPS = anterior intraparietal sulcus, LOTC = lateral occipitotemporal cortex, pIPS = posterior intraparietal sulcus, SPL = superior parietal lobule, PMd = dorsal premotor cortex, PMv = ventral premotor cortex.

correlated significantly with the group mean (all *ps* < 0.01, *one-sided*). Moreover, the MDS models of the two experiments strongly correlated with each other ($r(64) = 0.82, p < .001$, *one-sided*). This finding suggests that the MDS models reliably captured the subjective similarity across different types of stimuli, i.e., with and without real objects.

Subsequently, to test whether the participants' subjective similarity ratings were predicted by the TU model independent of the four control models (V, D, MD, and M), we conducted a multiple regression analysis. For experiment 1, the test revealed that the D model as well as the V model were no significant predictors (D: $\beta = 0.131$, $t(65) = 1.377$, $p = .174$; V: $\beta = 0.075$, $t(65) = 0.722$, $p = .473$). Therefore, these two predictor variables were deleted from the equation. Then, the remaining two models (MD and M) were included in the null model of the regression analysis. This test revealed that the null model explained 49.0 % of the variance in the subjective similarity ratings ($R^2 = 0.490$, $F(2,63) = 30.224$, $p < .001$) and the alternative model including the TU model as predictor variable explained additional 2.0 % of the variance ($R^2 = 0.510$, $F(1,62) = 2.563$, $p = .115$). The predictor variable of the TU model was no significant predictor of the participants' subjective similarity ratings, $\beta = 0.144$, $t(65) = 1.601$, $p = .115$. This means that the subjectively judged similarity barely changed for a higher TU similarity in experiment 1.

For experiment 2, we proceeded in the same way and found slightly different results: As in experiment 1, the initial test of the control models revealed that the D model as well as the V model were no significant predictors of the subjective similarity ratings (D: $\beta = 0.165$, $t(65) = 1.943$, $p = .057$; V: $\beta = 0.209$, $t(65) = 1.692$, $p = .096$). Therefore, these two predictor variables were deleted from the equation. Then, the MD model and the M model were included in the null model of the regression analysis. The test indicated that the null model explained 49.8 % of the variance in the subjective similarity ratings ($R^2 = 0.498$, $F(2,63) = 31.288$, $p < .001$) and the alternative model including the TU model as predictor variable explained additional 5.1 % of the variance ($R^2 = 0.549$, $F(1,62) = 6.943$, $p = .011$). Thus, the predictor variable of the TU model was a significant predictor of the participants' subjective similarity ratings in experiment 2, $\beta = 0.229$, $t(65) = 2.635$, $p = .011$. This means that the subjectively judged similarity increased for a higher TU-structure similarity in experiment 2.

In sum, a small portion of the variance in the participants' subjective similarity ratings was significantly explained by the TU model in experiment 2 only. In both experiments, the M model and the MD model were significant predictors of the similarity ratings.

3.2. Neuroimaging results

As expected, the searchlight GLM RSA revealed that in both, experiment 1 and experiment 2, the TU model predicted the representational organization of actions in the left anterior intraparietal sulcus (aIPS). The clusters in aIPS found in the two experiments strongly overlapped, peaking in the ventral postcentral gyrus and extending anteriorly into the central sulcus and posteriorly into the supramarginal gyrus (Fig. 3). Experiment 1 revealed an additional cluster for the TU model in the left occipital cortex.

The effect of the TU model in left aIPS indicated that the similarity in terms of TU sequence explained the representational variance in this area over and above the control models (for the searchlight result maps of the other models, see the Supplementary Material).

To test whether the aIPS differs significantly from other regions of the action observation network in this respect, we performed a GLM RSA in ROIs of the action observation network (based on univariate cluster peaks), which allows a more sensitive quantitative comparison of RSA effects in the different ROIs. This was done as it could be that also other regions of the action observation network are sensitive to the TU sequence, but failed to survive the conservative correction for multiple comparisons in the searchlight analysis.

Preparatory for the ROI RSA, to see whether the effects in the ROIs dissociate, a repeated measures ANOVA with between-subject factor Experiment(2), and within-subject factors ROI(11), and Model(6) was carried out. The Greenhouse-Geisser corrected results revealed significant main effects for ROI ($F(10,620) = 7.529$, $p < .0001$, $\eta^2_G = 0.006$) and Model ($F(3.39,210.07) = 19.919$, $p < .0001$, $\eta^2_G = 0.094$) as well as a significant interaction effect between ROI and Model ($F(17.78,1102.61) = 6.7$, $p < .0001$, $\eta^2_G = 0.060$) but no main effect for Experiment ($F(1,62) = 0.363$, $p = .549$, $\eta^2_G = 0.0002$). All remaining interaction effects were significant (all *ps* < 0.001).

The ROI RSA (Fig. 4) revealed that in the left aIPS, not only the TU

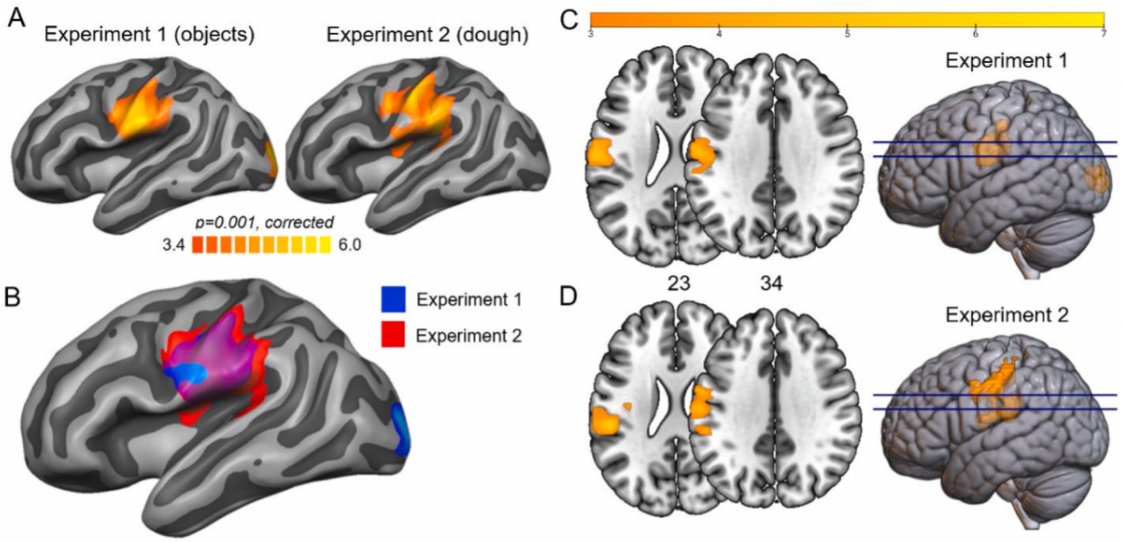

Fig. 3. Searchlight GLM RSA result for the TU model separately for experiment 1 and experiment 2 (A), and overlapping (B). Furthermore, (C) and (D) show the results for experiment 1 and 2, respectively, in axial slices (z coordinates given between slices). The peak coordinates are given in the Supplementary Material. Maps were thresholded using Monte Carlo correction for multiple comparisons.

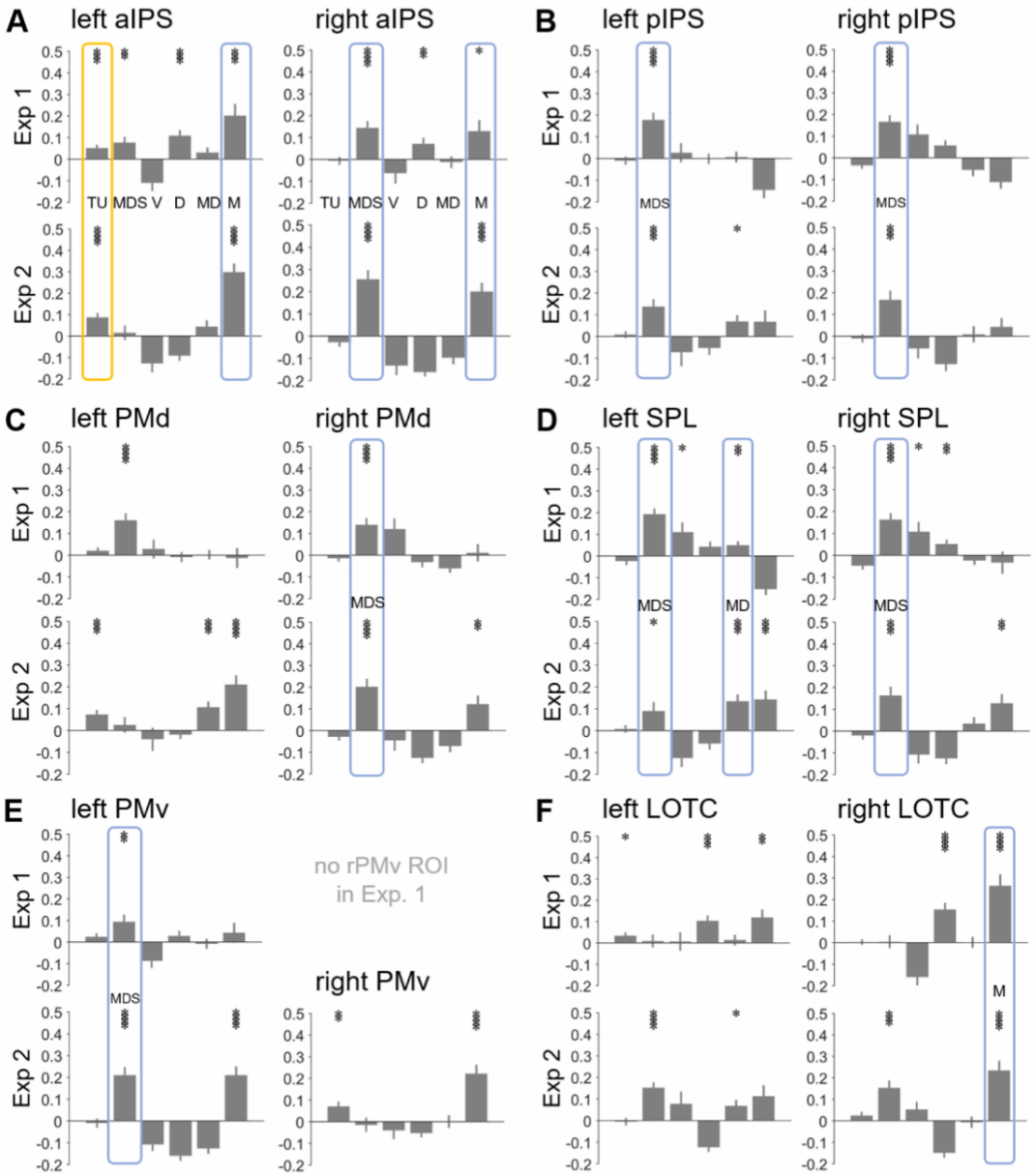

Fig. 4. ROI RSA results. Each bar chart shows the correlation of a neural RDM with the six model RDMs. A – F show the results for each bilateral ROI in experiment1 (Exp 1, top row) and experiment 2 (Exp 2, bottom row). For experiment 1, no right PMv ROI was defined as represented by blank space in the figure. Colored frames highlight matching results between experiment 1 and experiment 2. The yellow frame highlights common results regarding the TU model. Blue frames highlight matches for all the other models. Model abbreviations: TU - touching-untouching category model, MDS - inverse multidimensional scaling model, V - low-level visual model, D - duration model, MD - movement direction model, and M - manual model. aIPS = anterior intraparietal sulcus, pIPS = posterior intraparietal sulcus, PMd = dorsal premotor cortex, SPL = superior parietal lobule, PMv = ventral premotor cortex, LOTC = lateral occipitotemporal cortex. * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$, **** $p <= 0.0001$.

model explained the representational variance. In fact, most variance was explained by the M model. This did not come as a real surprise, since the aIPS is a critical region for object manipulation and tool use, which usually mainly draws on the dominant hand but often also requires the coordinated interplay of the dominant hand with the passive, stabilizing hand.

Furthermore, experiment 1 revealed additional effects in the left aIPS for the MDS model and the D model. Interestingly, the representational profiles in other areas clearly dissociated from that of the left aIPS. Thus, the right aIPS also revealed effects for the M model, but not for the TU model. Instead, we observed a clear effect for the MDS model. Interestingly, common effects were revealed for the MDS model in several ROIs: right aIPS, left PMv, bilateral pIPS, bilateral SPL, and right PMd. Right lateral occipitotemporal cortex (LOTC) revealed an effect for the M model and left SPL for the MD model in both experiments. In contrast, the left LOTC as well as the left PMd revealed no effects that were

consistent for both experiments.

4. Discussion

At a fundamental temporospatial level, object-directed actions can be described as touching-untouching (TU) sequences of hands, objects, and the ground. While actions that we observe can be classified in various ways, our study aimed to investigate whether our brain employs this TU structure as a sparse classification code for actions, and whether this code is reflected in explicit action categorization. Findings were consistent across two separate experiments involving either real objects or weakly-informative dough objects: Neuroimaging results revealed that the TU structure of actions is represented in the left anterior intraparietal sulcus, independent of other perceptual and semantic factors that may co-vary with this structure. At the behavioral level, we identified a subtle yet significant correlation between the TU structure and observers' judgments of action similarity with the dough objects. This suggests that the TU structure of actions can predict action categorization, particularly in contexts where object information is limited.

Just as technology is sometimes inspired by nature, the current brain activation study was inspired by robotics. Sequences of TUs are highly informative and useful for robots to recognize and execute object manipulations (Aksoy et al., 2011). Our findings suggest that they provide an organizing principle in human action recognition, too. There could be various reasons for this. For instance, TUs are particularly salient and thus easily recognizable incidents, since touching is always accompanied by deceleration and untouching announces acceleration of our movements. Accordingly, TUs are also particularly informative. At the same time, the recognition of TU sequences does not require the ability to distinguish between hands, objects, and the ground or to identify different objects. It therefore appears to be an ideal starting point for learning categories of action even before critical object expertise is built up (see Hunnius & Bekkering, 2010, for the early development of object knowledge). In this vein, it has been proposed that object-action association may develop earlier than object-word association (Eiteljoerge et al., 2019). Accordingly, preverbal infants might use TU sequences to more efficiently encode and more easily recognize everyday object manipulations they observe.

In our previous studies, we showed that the TUs in an object-manipulation action video are important and reliable anchor points for subjective event boundaries (Pomp et al., 2024, 2021). In addition to this behavioral relevance of TUs, we showed specific neural processing patterns differentiating between touchings and untouchings. Specifically, the difference in brain activity between touchings and untouchings suggested distinct cognitive processing roles: touchings strongly engaged visual regions, likely reflecting bottom-up visual processing, while untouchings recruited broader regions involved in updating expectations, highlighting the brain's response to action transitions and anticipation of what comes next.

The current results expand on these findings and show that TUs are not only meaningful for action segmentation but also serve as meaningful information in neural action category representation. Specifically, the left aIPS represents actions with similar TU structure similarly, which was shown at the whole-brain level and was unique in the examined action observation network. Generally, the aIPS is involved in creating an action plan for reach and grasp actions (see Turella & Lingnau, 2014, for a review), it codes hand and tool actions (Cabrera-Álvarez and Clayton, 2020), represents skills and conceptual action knowledge (Johnson-Frey, 2004), stores abstract representations of specific object-directed actions (Chen et al., 2018), and is involved in physical scene understanding (Fischer et al., 2016). The left aIPS has been described as one of the brain regions that are generally capable of discriminating actions of distinct categories and specifically of object manipulations (Wurm et al., 2017). Recent work by Wurm and Erigüç (2024) found the aIPL to encode abstract representations of cause-effect structures that capture the effect that is induced by an effector-target interaction in both observed human actions and abstract animations. Thus, our current finding that the TU structure of actions is represented in aIPS fits well the current state of research. Given the numerous similar characterizations of aIPS in terms of its role in hand-object interaction, TU sequences might even be a simple explanatory abstraction of this functionality.

Importantly, the present findings do not suggest that the left aIPS is specifically involved in the representation of touchings and untouchings per se. Rather, this area reflects the sequential pattern of touchings and untouchings that characterize different action categories. This is also supported by previous studies in which the contrast between touchings and non-TUs showed no engagement of the aIPS but significant activation in regions associated with visual processing, specifically the cuneus and lingual gyrus (Pomp et al., 2024, 2021). Engagement of these visual areas in response to touchings highlights their role in providing perceptual anchors for action segmentation in visual contexts. In accordance with this, the conjunction of untouchings contrasted to non-TUs in real-object actions and dough-object actions neither showed aIPS activation (Pomp et al., 2024).

In the same vein, our findings do not reflect differences in hand configuration. Previous studies showed the importance of grip similarity for ratings of manipulable objects (Hussain et al., 2024) as well as the influence of similarity in magnitude of arm movement and the hand configuration during use (Watson and Buxbaum, 2014). At first glance, one might assume that TU categories correspond to different grip types; however, this potential confound was ruled out, as TU categories in the present work did not systematically covary with specific hand postures.

Furthermore, to integrate our results into existing research, it is crucial to inspect the relationship between TU sequences and other formal descriptions of actions. A sequence of TUs describes an action on a very basic level. It captures the temporal dynamics of object manipulation—specifically, when and where contact happens or is lost. TU sequences might be a basic, foundational code that can be used to represent and recognize actions from early on and throughout life. Though TU sequences do not differentiate between agents or objects and can code contact states between item surfaces in object interactions without any agent involved (e.g., a branch that breaks off the tree in a storm and separates two apples lying on the ground is coded as destroy). Therefore, these sequences do not directly represent an abstract functional goal itself, but they form the underlying frame-work through which abstract functional goals like "rearranging" or "destroying" are realized and can correspond to them.

At the behavioral level, observers' similarity judgments of actions on minimally informative dough objects significantly related to the TU structure. While real objects might prompt action classification based on categorical associations, the reduced object information in the dough condition highlighted the TU structure, making it a key, informative basis for similarity judgments. This finding aligns with our previous study (Pomp et al., 2024), which showed that when object information is weak, subjective event boundaries fall systematically closer to touchings, emphasizing their perceptual relevance. Interestingly, comparing planning actions with unfamiliar in contrast to familiar objects, Van Elk et al. (2012) suggested a stronger goal-representation for familiar objects and a stronger motor imagery for unfamiliar objects, which points in the same direction as the current results. Also, violation paradigms where the object did not match the grip or goal of an action revealed independent neural temporal dynamics of the integration of motor acts and goal-related information during action observation (Decroix et al., 2020). In a similar vein, Bach et al. (2009) describe at least two partially distinct subprocesses involved in deriving both the motor act and the function of objects, which integrate during recognition into a unified action representation. To judge the similarity of the dough-object actions in the current study, participants may have preferably used the motor act information which is closer to TU structures than functional goals implied by familiar objects.

Remarkably, the MDS models derived from the two behavioral

similarity ratings accounted for a significant proportion of the variance in neural representational profiles across several regions in the action observation network of both experiments: bilateral posterior IPS, bilateral SPL, right aIPS, right PMd, and left PMv. This alignment suggests that behavioral similarity judgments resonate with the neural encoding of action structure, even in regions not directly tied to touching-untouching representation.

4.1. Limitations

In the multi-arrangement task, participants arranged triplets of single frame images that were supposed to remind them of the well-known action videos. While it is not particularly uncommon to use static images in action observation paradigms (cf. Caspers et al., 2010), future research might profit from presenting the entire action video in multi-arrangement tasks. Especially when participants are not as familiar with the actions as in the present work.

Furthermore, we applied a range of perceptual and semantic control models to confirm the specific, independent representation of TU structure. Future research could examine whether the TU model primarily reflects TU-based SEC categorization or if it aligns with additional perceptual and semantic principles beyond those assessed in this study. It could also profit from using more sophisticated visual control models than pixelwise similarity. Furthermore, it remains an open question whether the TU structure of an action plays a role in the recognition process or merely emerges as a byproduct of action categorization. This leads to the possibility that observers recognize actions by category, and because these categories covary with TU structure, we observe corresponding effects in both brain and behavior. Ultimately, these findings establish TU structure as a key element in how actions are perceived and recognized, opening new avenues for exploring the cognitive and neural bases of action segmentation.

4.2. Conclusion

Describing actions as an abstract sequence of relational changes in the form of touchings and untouchings of objects, hands, and the ground can effectively guide automated systems like robots in recognizing and responding to human actions without relying on object identification. The current study examined whether the neural processing of actions and their behavioral classification relies on the action categorization derived from their TU sequence. Using fMRI-based multivoxel pattern analysis, we identified neural representations of actions in the anterior intraparietal sulcus to be particularly associated to this TU structure. While models of one- or two-handedness and behavioral similarity ratings also explained representational activity in the action observation network, only the TU category model selectively correlated with the representational profile of the aIPS. These findings suggest that sequences of touchings and untouchings serve as a key organizing principle in inferior parietal action representation, extending prior research on meaningful anchor points for subjective event boundaries.

CRediT authorship contribution statement

Jennifer Pomp: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Moritz F. Wurm:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Formal analysis, Conceptualization. **Rosari N. Selvan:** Writing – review & editing, Validation, Formal analysis. **Florentin Wörgötter:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Ricarda I. Schubotz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would especially like to thank Jasper J. F. van den Bosch for his great help in implementing the multi-arrangement task on the meadows research platform. Furthermore, we would like to thank Minija Tamosiunaite and Tomas Kulvicius for their help with the methods setup. Finally, we thank Monika Mertens, Theresa Eckes, Katharina Thiel, Alina Eisele, Mina-Lilly Shibata, Simon Reich, Yuyi Xu, and Annika Garlichs for their assistance during stimulus material creation or data collection.

Funding

This work was supported by the German Research Foundation (DFG) [grant numbers SCHU 1439/8–1, WO 388/13–1]. The DFG had no involvement in study design, data collection, analysis and interpretation of the data, writing of the report, or decision to submit for publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2025.121113.

Data availability

The multi-arrangement task stimuli, iMDS data as well as the ROI data have been deposited in an OSF repository (DOI 10.17605/OSF.IO/9V3CQ). The video stimulus material is available via the Action Video Corpus Münster (AVICOM, https://www.uni-muenster.de/IVV5PSY/AvicomSrv/). The raw data of the behavioral multi-arrangement task and the fMRI analyses is available upon reasonable request.

References

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

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, 1229–1249. https://doi.org/10.1177/0278364911410459.

Bach, P., Gunter, T.C., Knoblich, G., Prinz, W., Friederici, A.D., 2009. N400-like negativities in action perception reflect the activation of two components of an action representation. Soc. Neurosci. 4, 212–232. https://doi.org/10.1080/17470910802362546.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188. https://www.jstor.org/stable/2674075.

Buxbaum, L.J., Randerath, J., 2018. Limb apraxia and the left parietal lobe. Handbook of Clinical Neurology, 1st ed. Elsevier B.V. https://doi.org/10.1016/B978-0-444-63622-5.00017-6

Cabrera-Álvarez, M.J., Clayton, N.S., 2020. Neural processes underlying tool use in humans, macaques, and corvids. Front. Psychol. 11, 1–11. https://doi.org/10.3389/fpsyg.2020.560669.

Caspers, S., Zilles, K., Laird, A.R., Eickhoff, S.B., 2010. ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50, 1148–1167. https://doi.org/10.1016/j.neuroimage.2009.12.112.

Chan, A.W.Y., Baker, C.I., 2015. Seeing is not feeling: posterior parietal but not somatosensory cortex engagement during touch observation. J. Neurosci. 35, 1468–1480. https://doi.org/10.1523/JNEUROSCI.3621-14.2015.

Chen, Q., Garcea, F.E., Jacobs, R.A., Mahon, B.Z., 2018. Abstract representations of object-directed action in the left inferior parietal lobule. Cereb. Cortex 28, 2162–2174. https://doi.org/10.1093/cercor/bhx120.

de Vries, I.E.J., Wurm, M.F., 2023. Predictive neural representations of naturalistic dynamic input. Nat. Commun. 14. https://doi.org/10.1038/s41467-023-39355-y.

Decroix, J., Roger, C., Kalénine, S., 2020. Neural dynamics of grip and goal integration during the processing of others' actions with objects: an ERP study. Sci. Rep. 10, 1–11. https://doi.org/10.1038/s41598-020-61963-7.

Eiteljoerge, S.F.V., Adam, M., Elsner, B., Mani, N., 2019. Word-object and action-object association learning across early development. PLoS One 14, 1–22. https://doi.org/10.1371/journal.pone.0220317.

Fischer, J., Mikhael, J.G., Tenenbaum, J.B., Kanwisher, N., 2016. Functional neuroanatomy of intuitive physical inference. Proc. Natl. Acad. Sci. U. S. A. 113, E5072–E5081. https://doi.org/10.1073/pnas.1610344113.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. Neuroimage 7, 30–40. https://doi.org/10.1006/nimg.1997.0306.

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, 189–210. https://doi.org/10.1002/hbm.460020402.

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, 1–4. https://doi.org/10.3389/fninf.2012.00022.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.R., Yuan, J., Zhao, W.X., Zhu, J., 2021. Pre-trained models: past, present and future. AI Open 2, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image Vis. Comput. 60, 4–21. https://doi.org/10.1016/j.imavis.2017.01.010.

Hunnius, S., Bekkering, H., 2010. The early development of object knowledge: a study of infants' Visual anticipations during action observation. Dev. Psychol. 46, 446–454. https://doi.org/10.1037/a0016543.

Hussain, A., Walbrin, J., Tochadse, M., Almeida, J., 2024. Primary manipulation knowledge of objects is associated with the functional coupling of pMTG and aIPS. Neuropsychologia 205, 109034. https://doi.org/10.1016/j.neuropsychologia.2024.109034.

JASP Team, 2024. JASP (Version 0.18.3) [Computer software].

Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., Tang, X., 2022. New generation deep learning for video object detection: A survey. IEEE Trans. Neural Networks Learn. Syst. 33, 3195–3215. https://doi.org/10.1109/TNNLS.2021.3053249.

Johnson-Frey, S.H., 2004. The neural bases of complex tool use in humans. Trends Cogn. Sci. 8, 71–78. https://doi.org/10.1016/j.tics.2003.12.002.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. https://doi.org/10.1073/pnas.0600244103.

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

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28. https://doi.org/10.3389/neuro.06.004.2008.

Liang, Q.A., Wan, T.F., 2020. Difficulty Within Deep Learning Object-Recognition Due to Object Variance. Springer International Publishing, pp. 278–289. https://doi.org/10.1007/978-3-030-63830-6_24.

Murata, A., Wen, W., Asama, H., 2016. The body and objects represented in the ventral stream of the parieto-premotor network. Neurosci. Res. 104, 4–15. https://doi.org/10.1016/j.neures.2015.10.010.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10. https://doi.org/10.1371/journal.pcbi.1003553.

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

Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in matlab/GNU octave. Front. Neuroinform. 10, 1–27. https://doi.org/10.3389/fninf.2016.00027.

Peelen, M.V., Caramazza, A., 2012. Conceptual object representations in human anterior temporal cortex. J. Neurosci. 32, 15728–15736. https://doi.org/10.1523/JNEUROSCI.1953-12.2012.

Pomp, J., Garlichs, A., Kulvicius, T., Tamosiunaite, M., Wurm, M.F., Zahedi, A., Wörgötter, F., Schubotz, R.I., 2024. Action segmentation in the brain: the role of object–Action associations. J. Cogn. Neurosci. 36, 1784–1806. https://doi.org/10.1162/jocn_a_02210.

Pomp, J., Heins, N., Trempler, I., Kulvicius, T., Tamosiunaite, M., Mecklenbrauck, F., Wurm, M.F., Wörgötter, F., Schubotz, R.I., 2021. Touching events predict human action segmentation in brain and behavior. Neuroimage 243, 118534. https://doi.org/10.1016/j.neuroimage.2021.118534.

Reynaud, E., Lesourd, M., Navarro, J., Osiurak, F., 2016. On the neurocognitive origins of human tool use: A critical review of neuroimaging data. Neurosci. Biobehav. Rev. 64, 421–437. https://doi.org/10.1016/j.neubiorev.2016.03.009.

Turella, L., Lingnau, A., 2014. Neural correlates of grasping. Front. Hum. Neurosci. 8, 1–8. https://doi.org/10.3389/fnhum.2014.00686.

Urgen, B.A., Pehlivan, S., Saygin, A.P., 2019. Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. Neuropsychologia 127, 35–47. https://doi.org/10.1016/j.neuropsychologia.2019.02.006.

Van Elk, M., Viswanathan, S., Van Schie, H.T., Bekkering, H., Grafton, S.T., 2012. Pouring or chilling a bottle of wine: an fMRI study on the prospective planning of object-directed actions. Exp. Brain Res. 218, 189–200. https://doi.org/10.1007/s00221-012-3016-9.

Vingerhoets, G., 2014. Contribution of the posterior parietal cortex in reaching, grasping, and using objects and tools. Front. Psychol. 5, 1–17. https://doi.org/10.3389/fpsyg.2014.00151.

Watson, C.E., Buxbaum, L.J., 2014. Uncovering the architecture of action semantics. J. Exp. Psychol. Hum. Percept. Perform. 40, 1832–1848. https://doi.org/10.1037/a0037449.

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, 117–134. https://doi.org/10.1109/TAMD.2012.2232291.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—Again. Neuroimage 2, 173–181. https://doi.org/10.1006/nimg.1995.1023.

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

Wurm, M.F., Erigüç, D.Y., 2024. Decoding the physics of observed actions in the human brain. BioRxiv 2023.10.04.560860. https://doi.org/10.1101/2023.10.04.560860.

Wurm, M.F., Lingnau, A., 2015. Decoding actions at different levels of abstraction. J. Neurosci. 35, 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15.2015.

J. Pomp et al. NeuroImage (2025)
Touching-untouching patterns organize action representation in the inferior parietal cortex

Supplementary material

Figures

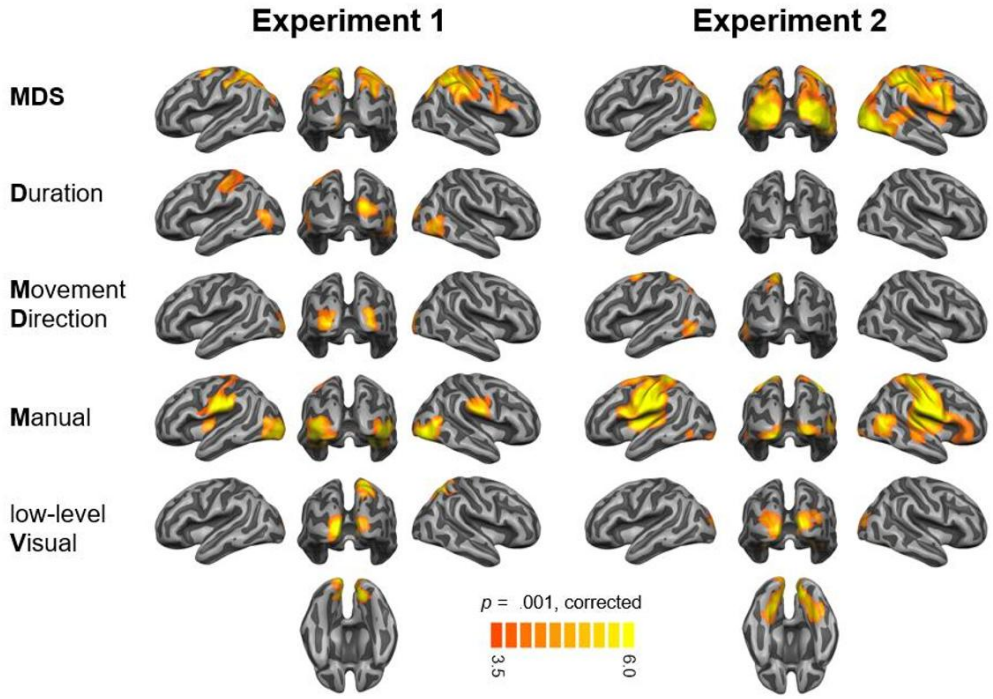

Fig. S1. Searchlight GLM RSA results for experiment 1 (left) and experiment 2 (right). Respective model names are given in the left column. Maps were thresholded using Monte Carlo correction for multiple comparisons. MDS = multidimensional scaling.

Tables

Table S1. Maxima of Activation from the Searchlight GLM RSA of the TU Model

Macroanatomical location	Abbreviation	H	Cluster Extent	*t*-value	MNI Coordinates		
					x	y	z
Experiment 1 (objects)							
postcentral gyrus	PCG	L	324	5.69	-63	-7	32
anterior intraparietal sulcus	aIPS	L		4.89	-51	-22	29
anterior supramarginal gyrus	aSMG	L		3.79	-60	-28	32
inferior occipital gyrus	IOG	L	143	7.63	-24	-100	2
Experiment 2 (dough)							
postcentral gyrus	PCG	L	557	5.80	-60	-16	41
parietal operculum	OPRpr	L		5.77	-51	-28	23
Posterior Insula	INS	L		5.24	-36	-10	20

Note. H = Hemisphere, L = Left.

4 General Discussion and Future Directions

The present work aimed to investigate event perception through subjective and objective event boundaries, their mutual relationship, underlying neural processing, emerging action categories, and their variability due to fluctuating object-action associations. Event perception serves as a central cognitive mechanism, structuring time in a way that enables the brain to form memories and generate predictions. Accordingly, it is essential to investigate this phenomenon to better understand its underlying mechanisms. To this end, we related observer-labeled boundaries to stimulus-derived boundaries, identified categories and both the boundaries and categories were used to model participants' brain activity. In two experiments, participants passively observed object manipulations of either commonplace items or dough items during the MRI scan. Subsequently, the manipulation videos were behaviorally segmented into events and finally spatially arranged to derive action categories from these similarity judgments via inverse multidimensional scaling. Objective boundaries and corresponding action categories were extracted using computer vision algorithms based on low-level stimulus features. To examine neural activation patterns, we applied univariate as well as representational similarity analyses to the fMRI data. In the following sections, the results and limitations of the experiments will be summarized and discussed.

4.1 Event Boundaries

The results of Experiment 1 and Experiment 2 revealed that subjective event boundaries did not match objective event boundaries but co-occurred systematically. Due to the different occurrence frequencies, it was still possible to disentangle the time-locked neural activity patterns. The following sections summarize and discuss the boundary-evoked brain responses,

the systematic temporal co-occurrence of different boundary types and the detection reliability to finally answer the first research question[4] in the conclusion of this section.

4.1.1 Boundary-Evoked Brain Responses

In the following, I will discuss the neural activation patterns underlying the different types of event boundaries that were mutually found in Experiment 1 and Experiment 2, as analyzed in Study II. Thus, these activation patterns are constant across different object types.

4.1.1.1 Brain Responses at Subjective Event Boundaries

Experiment 1 and Experiment 2 revealed increased bilateral brain activity at observer-labeled event boundaries in the lateral occipital cortex (LOC), posterior middle temporal gyrus (pMTG) and superior parietal lobule (SPL). Regarding the former, boundary-evoked activity at the occipitotemporal junction, spanning the motion-sensitive MT complex, has been described earlier (Schubotz et al., 2012; Speer et al., 2003; Zacks, Braver, et al., 2001; Zacks, Swallow, et al., 2006). The hypothesis that distinctive movement features matter in event structure perception was put forward early on (Newtson et al., 1977) and the current results further support the idea that event structure perception is related to the detection of visual change such as, for instance, changes in motion. Zacks (2004) found that the probability of identifying an event boundary was related to movement features so that observers have a tendency to mark event boundaries when objects are moving quickly. Furthermore, Zacks, Swallow, et al. (2006) showed that brain regions that process general and biological motion selectively respond at event boundaries. In fact, the brain activation patterns of the conjunction of boundary-evoked activation between Experiment 1 and 2 (as shown in Fig. 5 of Study II) include the mean location coordinates reported in Zacks, Swallow, et al. (2006; Table 2). Thus, the current results

[4]As previously introduced, the first research question reads:
Are TUs as objective event boundaries a meaningful supplement or reference point to subjective event boundaries and how do they contribute to understanding neural event structure processing in object-directed action observation?

replicated boundary-sensitivity in regions processing general motion and biological motion, but the question of the driving processes remains unanswered. It is possible that greater motion changes appeared at event boundaries and directly triggered bottom-up MT complex activation. On the other hand, the motion changes could have been predicted by higher-level cognitive processes (e.g., event models) that activate the MT complex top-down. Future research is needed to further elucidate the role of motion processing in event structure perception.

In addition to the posterior temporal and lateral occipital cortex, the SPL was found active. This region has been shown to be activated by controlling goal-directed limb movements and especially by reaching (Gamberini et al., 2020), and reach-to-grasp action (Fattori et al., 2017). Furthermore, it was frequently reported as being part of the action observation network (Hardwick et al., 2018). It has been found active when observing reaching to and grasping of objects and Wurm et al. (2017) suggested its activation to be related to body part motion in space. In the context of event structure perception, boundary-evoked SPL activation has barely been reported before. However, in the current work, the SPL activity could reflect the stimulus material showing reaching and grasping actions. Though, the precise nature of its role in time-locked subjective boundary processing remains elusive. For a better understanding of SPL's boundary sensitivity in object-directed actions, future research is needed. The current results suggest that motion and especially motion of the hand are essential for subjective event structure perception.

4.1.1.2 Brain Responses at Objective Event Boundaries: Untouchings

In contrast to subjective event boundaries, the processing of objective event boundaries has barely been investigated with functional imaging. It is remarkable that the simultaneous modeling of fMRI data with subjective and objective event boundaries split the pattern found so far for observer-labeled boundaries. In this section, I will address the results for objective untouchings (i.e., the point when two touching objects un-touch). While in Experiment 1 (as

reported in Study I) a more distributed pattern was found for untouchings, the results of Experiment 2 replicated only two cortical regions. Specifically, the conjunction of the brain activation to untouchings between Experiment 1 and Experiment 2 (as reported in Study II) revealed bilateral parahippocampal cortex and left dorsal premotor cortex (PMd) activity. Regarding the latter, dorsal premotor activity has been reported before for observing transitive versus intransitive actions (Wurm, Caramazza, et al., 2017), and it has been found for observer-labeled boundaries in object-directed actions (Schubotz et al., 2012). Furthermore, dorsal premotor (or caudal superior frontal sulcus) activity was found to be elicited by updating the attentional focus, together with the posterior parietal cortex (Bledowski et al., 2009). It has been suggested that PMd is specifically involved when the position of an object in space drives the spatial parameters of arm movements (i.e., reaching) given the spatial properties form the attentional focus (Schubotz & von Cramon, 2001). Furthermore, and important in the current context, left PMd has been shown to be involved in initiating action prediction during the observation of everyday actions (W. Stadler et al., 2011, 2012). Though this region has been reported rather rarely in the context of event segmentation, the opposite is the case for the PHC (see Baldassano et al., 2017; Reagh et al., 2020; Schubotz et al., 2012).

Boundary-evoked PHC activity has been shown in different age groups, and a decrease during aging has been reported (Reagh et al., 2020). Furthermore, Baldassano et al. (2017) reported on cortical event boundaries in PHC, that strongly related to hippocampal activity, suggesting that the hippocampus encodes information about the just-ended event into episodic memory and De Soares et al. (2024) demonstrated that the timing of cortical event boundaries in PHC are influenced top-down. Besides, PHC activity can reliably be seen when contextual associative information is encoded or retrieved from memory (Aminoff et al., 2013; Li et al., 2016), especially in spatial and episodic memory (Geva-Sagiv et al., 2023). In sum, the current state of research is far from providing a clear picture of the PHC involvement in event structure processing though its involvement as such is not in question. Concurrent activation of PMd and

PHC in action observation has been found for action prediction (W. Stadler et al., 2011) and for action boundary detection (Schubotz et al., 2012). Thus, the current results suggest that at untouchings the attentional focus is updated and the end of an event is signaled so that it may be encoded in the hippocampus. Furthermore, the found activation pattern indicates the prediction of the upcoming action step.

4.1.1.3 Brain Responses at Objective Event Boundaries: Touchings

Touching incidents could also be clearly separated from untouchings and observer-labeled event boundaries regarding their corresponding brain response. The conjunction of the brain activation to touchings between Experiment 1 and Experiment 2 (as reported in Study II) revealed increased bilateral cuneal and lingual gyrus activation. These results are consistent with past studies showing that the activation of medial occipital areas close to the calcarine sulcus reflect low-level visual differences between stimuli in the representation of local scene elements (Kamps et al., 2016). In addition, increased cuneus and lingual gyrus responses were shown for allocentric versus egocentric spatial representations (Ruotolo et al., 2019). The medial occipital lobe is a highly interconnected system that performs coordinated basic visual processing and has many long-range association fibers supporting language and memory functions (Palejwala et al., 2021). Increased medial occipital lobe activity in response to the emerging surface contact (i.e., touching) between two objects points to increased visual inspection of the scene. This information may then be propagated to align with the prediction of the subsequent action step.

4.1.2 Co-Occurrence of Objective and Subjective Event Boundaries

The results of Experiment 1 and 2 revealed a systematic co-occurrence of group-determined event boundaries and objective (un)touchings. Their distribution over time showed clear peaks of subjective boundaries shortly after a touching relation emerged and a more

widely distributed emergence of subjective boundaries around untouchings. Subjective boundaries were significantly closer related to TUs than random button presses. Hence, some touching and untouching incidents were important anchor points for behavioral action segmentation. The T-U motif and video content analyses of Experiment 1 (as reported in Study I) further revealed that subjective boundaries especially marked certain action phases, that is, during the object manipulation and at the onset of object transport. During hand transport, during object transport, and at the end of object transport, in contrast, subjective boundaries were less prevalent.

The current results represent an initial step toward revealing objective anchors for subjective boundary markings. They suggest that at least some points of touching and untouching are relevant to predict participants' segmentation behavior, though many are not. Strikingly, the mere frequency of the incidents varies greatly so that further identification and characterization of relevant TU incidents will be necessary. To this aim, object identity seems an obvious candidate to specify TUs. A current approach that specifically examined hand-object touchings and untouchings (ignoring the contact states between objects, and object to ground), for instance, turned out to be useful in marking action steps in action prediction processing (Selvan et al., 2024). Considering these results, it may be the touching between the hand and the object that primarily gave rise to subjective event boundaries in the current experiments. The onset of the object transport was coded as the phase between the touching of the hand with the object and the subsequent untouching of the object from the ground. Here, especially the emerging hand-object contact may have been crucial. Regarding the second important action phase (i.e., during the object manipulation), it was coded as the period between the timepoint when the hand touches the object to manipulate and when the hand un-touches from this object after manipulation. Here again, the emerging hand-object contact may have been decisive.

It has been argued that the sequence of touching and untouching between objects (including the hand and the ground) may be a fruitful source of information for preverbal infants. Nonetheless, it cannot be denied that during development we learn to distinguish hands from objects in the sense that hands are the effectors of an agent and therefore different from inanimate objects. In fact, anticipatory eye movements have been demonstrated for hand-object interactions at the age of 12 months (Falck-Ytter et al., 2006) and in adults (Flanagan & Johansson, 2003), and when the possibility existed to infer an autonomous agent as causing the observed movements (Gesierich et al., 2008). It would therefore make sense if the greater relevance of the acting effector is reflected in a higher relevance of its contact with manipulable objects. This hypothesis has yet to be tested. In order to do this, it may even be possible to selectively reanalyze the data of this thesis. Specifically, the changes in contact states between the hand and a manipulable object could be identified and selected to repeat the analyses with this part of the data regarding their temporal co-occurrence to subjective event boundaries. This could offer a preliminary insight and inspire further studies. According to the present status, our results reveal a significant role of objective event boundaries for subjective event structure perception that lies the groundwork for further, in-depth research.

4.1.3 Reliability in Event Boundary Detection

Previous research repeatedly demonstrated the high intra- and interindividual reliability in behavioral action segmentation (for a review see Sasmita & Swallow, 2023). Correspondingly, the retest reliability results from Experiment 1 and Experiment 2 confirmed consistent unit marking behavior on the individual and on the group level. Especially the comparison of the behavioral segmentation data to simulated random button presses, that preserved the stochastic characteristics of the individual behavior, validated that the segmentation followed a nonrandom pattern. In sum, these results confirm that the unit marking procedure is a valuable

and useful method to determine subjective event boundaries. The comparison to simulated button presses complements the reliability check meaningfully.

The objective event boundaries were determined by computer vision. Automated visual action recognition is a rapidly growing field, and the methods applied depend inter alia on the data's input format. In the current work, RBG-D data was used meaning that the algorithm utilized depth images to generate point clouds in the first step of automated TU time point extraction. Compared to manual time point extraction, which would need a time-consuming frame-by-frame inspection of each action manipulation video by several raters, the algorithm bears a significant efficiency gain. Moreover, it successfully detects emerging and disappearing touching relations. However, the algorithm does not guarantee perfect performance. Occasionally, it misinterpreted the scene and needed manual correction. As stated above, it used depth images which are essentially two-dimensional data while three-dimensional data would probably improve its performance considerably. Especially the precise determination of an emergence of a touching relation and its disappearance can depend on the perspective on the scene. Viewed from an unfavorable perspective, the contact between surfaces can be occluded. Fortunately, the rapid development of virtual reality (VR) techniques that goes along with high fidelity three-dimensional camera setups promises even better input data quality and visual action recognition and segmentation algorithms to come.

4.1.4 Understanding Event Structure Perception Through Objective Boundaries

One of the aims of this work was to determine whether objective event boundaries meaningfully supplement subjective event boundaries and how they contribute to understanding (neural) event structure perception. Concerning segmentation behavior, objective event boundaries seem to be valuable anchor points for subjective event boundaries. Concerning the neural processing of event structure, the activation patterns underlying

subjective and objective event boundaries were clearly distinguishable and could be functionally interpreted. Observer-labeled event boundaries were mainly accompanied by increased motion processing. This pattern is consistent with previous findings and suggests that participants tend to segment actions dependent on motion features. Previously identified regions beyond this could now be assigned to objective event boundaries. Despite behavioral data implying that the moment of touch could be the most critical one, this is not the case. Medial occipital activation patterns appear to reflect merely increased visual inspection of the scene when objects touch, rather than more complex processing. In contrast, the moment when a touching relation is released (i.e., the untouching) marks the point at which attention is redirected, the completion of an event is signaled to memory and upcoming action is predicted.

In sum, it can be concluded that objective event boundaries constitute a meaningful addition to subjective event boundaries to investigate event structure perception in object-directed actions. Fortunately, objective event boundaries offer several advantages as they can be determined a priori independently from the participants' perception, and they can deliberately be manipulated to a certain extent to design experiments. Eventually, according to the current state of research, when investigating the neural processing of event structure, objective event boundaries can be seen as a supplement to subjective boundaries rather than a replacement.

4.2 When Objects Suggest Action

Objects that we use in our daily life carry a lot of information about what we can do with them (El-Sourani et al., 2018, 2019; Hrkać et al., 2015; Kalénine et al., 2016; Schubotz et al., 2014) and the strength of these object-action associations varied between experiments. In Experiment 1, the associations were strong, whereas in Experiment 2, they were weak or nonexistent. This is because in Experiment 1 commonplace items were manipulated in the

action videos and in Experiment 2 these items were replaced by formed pieces of dough. The idea was motivated by the fact that the employed computer vision algorithm does not identify objects, nor does it use object-related knowledge. While adult humans benefit from a life-long experience with manipulable objects, the algorithm could be a model for early infants' action perception. Furthermore, the time-locked analyses of objective event boundaries that sparsely code movement information allowed the dissociation between movement effects and object effects. Remarkably, the difference in association strength had no major effect on the neural processing of objective touchings and of observer-labeled event boundaries, but even more so when it comes to untouchings and segmentation behavior. The following sections summarize and discuss the effects of strong and weak object-action associations separately to finally address the second research question[5] in this section's conclusion.

4.2.1 Strong Object-Action Associations

The neural activity increase due to strong object-action associations was relatively small. Concerning the entire action duration, objects with strong action associations evoked increased activity in right anterior supramarginal gyrus (aSMG) and anterior intraparietal sulcus/ ventral postcentral sulcus with ROI results being significant in bilateral aIPL. This is consistent with previous research by Meyer et al. (2011) who showed 5-second videos of bimanual object exploration to their participants and revealed that the most discriminative voxels (i.e., when discriminating between videos) were located in the postcentral sulcus and on the posterior wall of the postcentral gyrus with a right hemispheric dominance. In a follow-up study, Kaplan and Meyer (2012) extended these results and inferred that stimulus-specific patterns of activity around the intraparietal sulcus bear high information content. While the authors interpreted their results with respect to somatosensory processing of haptically perceived shape of different

[5] As previously introduced, the second research question reads:
Do object-action associations, provided by the manipulated object, modulate action segmentation behavior and the neural processing at subjective and objective event boundaries?

objects, this is not the only possible interpretation. Each video showed one everyday object that was haptically explored. Therefore, the observed brain activation pattern could also be interpreted as it was proposed by Schubotz et al. (2014). They showed that activity in the aIPL varied as a function of the number of actions that participants associated with objects. Similarly, Wurm and Schubotz (2018) contrasted naturalistic versus pixelized object-directed actions and found stronger neural responses in bilateral postcentral gyrus extending into aSMG. Both naturalistic and pixelized stimuli showed the kinematics of the action, but only the naturalistic one allowed recognition of the object and to use the related information. Remarkably, increased right ventral postcentral gyrus activation for grasping an everyday object versus grasping a geometrical shape was even found for action imagination (Schulz et al., 2018). Thus, the precise nature of aIPL's role is currently the focus of research and object identity including associated information could be pivotal, as in the case of strong object-action associations.

Concerning the time point specific analyses, the effect of strong object-action associations became apparent at untouchings. A particular role of mnemonic associations in the presence of strong object-action associations was hypothesized and was indeed reflected in increased parahippocampal responses. This result was consistent with previous findings in which the parahippocampal cortex has reliably been reported for the encoding and retrieval of contextual associations (Aminoff et al., 2013). Thus, when robust object-action associations were available, this information was retrieved from memory to segment and predict actions. Furthermore, increased cuneal and lingual activity was detected at untouchings which indicated increased visual inspection, as discussed earlier. This could be due to commonplace items being visually more detailed than formed pieces of dough. The visual information could then be used to identify the object and predict the unfolding action.

4.2.2 Weak Object-Action Associations

The neural activity increase due to weak object-action associations was surprisingly large. At untouchings, our hypothesis was corroborated regarding an increase in activation in biological motion processing areas. Thus, without strong object-related predictions for the upcoming action, motion processing gained importance for segmenting and predicting object-directed actions. This may be frequently the case for infants, for whom it is perfectly normal to encounter unknown objects, and rarer for adults. As early as at the age of four months, infants show a preference for biological motion patterns (Fox & McDaniel, 1982) and functional object knowledge of familiar objects solidifies shortly afterwards (Hunnius & Bekkering, 2010).

Furthermore, weak object-action associations have had a significant effect on the segmentation behavior. The behavioral measures of reliability and consistency, as well as unit marking frequencies, and systematicity were broadly comparable between Experiment 1 and 2, but the group retest reliability and behavioral systematicity were higher in Experiment 2 going along with a smaller variance in segmentation behavior. Thus, participants set event boundaries more often and closer to touching events when objects were weakly informative. The occurrence of a touching is frequently associated with a reduction in speed, eventually coming to a stop. Here, this acted as a reference to segment the action. When the object did not offer specific information about what action to expect, this movement sequence gained importance. The behavioral results thus paint the same picture as the above-mentioned increase in brain activity.

Unexpectedly, weak associations yielded increased aIPL activation at untouchings. Based on previous studies, showing increasing aIPL activity with an increasing number of correlated actions (Schubotz et al., 2014), we rather expected an increase for strong associations, which could indeed be seen when analyzing the entire duration of the action and has been discussed above. In contrast, the time-locked effect at untouchings was interpreted as reflecting an unrestricted number of candidate actions in the case of dough manipulations. This

is, object-action associations restrict which actions are considered to unfold next and in the case of dough items, the search space for candidate actions is unrestricted by these associations. A comparable effect has been described before. Wurm and Schubotz (2012) showed that the pre-activation of expectable actions by contextual cues reduced the search space for input-to-memory matching by biasing those actions that are most probable in a given context. In a subsequent study, Wurm, Artemenko, et al. (2017) examined contextual factors in children between four and eight years of age and revealed that they effectively integrated contextual information in action recognition and profited the most from context information when actions were unfamiliar. Apparently, the aIPL's functional profile is many-faceted and I will discuss the role of aIPL comprehensively in section *4.4 The aIPL in Action*.

4.2.3 Action Processing Modulated by Object-Action Associations

One of the aims of this work was to determine the effect of object-action associations on action segmentation and its neural processing. The current results confirm that object-related knowledge modifies how object-directed actions are processed. Segmentation behavior becomes more targeted to objective event boundaries when objects do not offer rich information about what to expect and movement information gains importance when the prediction of the next step is synchronized with the perceptual input. Moreover, at this point of synchronization, increased aIPL activation indicated an unrestricted search for candidate actions when object information was limited. While this latter effect was time-point specific, the contrary was the case across the whole period of the action. Concerning strong action associations, the rich visual information that commonplace objects offer is crucial for the action step synchronization.

Despite these modulations, there were major similarities for observing manipulations of objects that were either strongly- or weakly-associated with actions. The conjunction analyses (Study II) showed a large common pattern of brain activity that replicated the well-established

action observation network that I thoroughly described in section *1.2.1 The Action Observation Network*. It showed a large cluster in the posterior temporal and lateral occipital cortex that ventrally extended in the parahippocampal cortex and hippocampus. This activity was supplemented by a large parietal cluster that spanned along the postcentral sulcus in parallel to the precentral sulcus activation that in turn extended into insular regions. Thus, object information makes a notable difference in object-directed action processing though common patterns clearly predominate.

4.3 Action Categories

It is one of the current challenges of cognitive neuroscience to understand semantic representation in the brain. Previous research has shown patterns for various contents (for a review see Binder et al., 2009), though these representations are dynamic and, for instance, are modulated by attention (Çukur et al., 2013) with results depending on methodological choices in multivariate imaging (Frisby et al., 2023). Nonetheless, in the current work, we aimed to examine the representation of action categories. Actions can be grouped in categories according to their patterns of touching and untouching. At the same time, actions can also be classified according to participants' judgments of similarity. The following sections discuss how these categorizations are represented in the brain and how they relate to one another to address the third research question[6].

4.3.1 Objective Action Categories

Wörgötter et al. (2013) showed that action categories can be derived from TU sequences. The current work showed in study III that these action categories are associated with the neural

[6] As previously introduced, the third research question reads:
Are TU action categories represented in neural processing patterns of object-directed actions and are they related to behavioral action classifications?

representations in the left aIPS. These cross-experiment results highlight the value of objective TU sequences to understand neural action representation. Regarding the functional profile of the aIPS, previous action observation research employing repetition suppression has suggested that the left aIPS is particularly important for processing the identity and function of a grasped object, independent of grip and trajectory, and particularly involving goal representations (Grafton & Hamilton, 2007).

It could be argued that TU action categories covary with abstract functional action goals. This would mean that turning and pulling an object share their TU sequence and both achieve the functional goal of rearranging. Theoretically, TU sequences code the contact states between items irrespective of their identity or agency and accordingly, they cannot directly represent abstract functional goals but only offer an underlying framework through which abstract functional goals are realized. Practically, the SEC framework that has originated TU sequences was developed to let robots learn the semantics of object action by observing humans' object-directed actions (Aksoy et al., 2011) which means that agency cannot be ignored. Accordingly, the stimulus material employed in the experiments displayed TU sequences embedded in action videos with an agent and a goal. Thus, the activation of the aIPS could mirror the covarying goals.

This would align with previous findings, suggesting the aIPS/aSMG to represent the abstract function or purpose of an action (i.e., decorating or protecting; Leshinskaya & Caramazza, 2015). Similarly, Urgen et al.'s (2016) RSA results showed that the models for category of action, intention of action, and target of action, all correlated best with the parietal node of the action observation network. However, these considerations are speculative and functional goals or intentions were not the primary focus of this work. It should be investigated more systematically in future studies.

To this end, one possibility would be to create stimuli that share the same TU sequence but have no functional goal. It may prove to be difficult to design reasonable videos without an agent, though, as the TU sequence specifies that some object manipulates another object. Even if the first object is not as obvious an agent as the hand, participants would most likely read it as the agent and attribute goals. Previous research has shown that already three-month-old infants attribute goals to the actions of novel non-human agents (Luo, 2011). Accordingly, there are different interpretations of the current results that emphasize different functional profiles of the aIPL. It must be noted, however, that TU-based action categories being represented in the aIPL is already a remarkable finding in itself, and a better understanding of these representations is an important future research aim.

4.3.2 Subjective Action Categories

Using inverse multidimensional scaling served to gather subjective action categories, respectively a subjective action space. The ROI RSA results of Study III revealed that the MDS models explained a significant part of the representational variance in several brain regions of the action observation network across experiments, i.e., the right aIPS, left PMv, bilateral pIPS, bilateral SPL, and right PMd. Thus, a large proportion of the action observation network was associated with the behavioral action classifications. The question arises which dimensions participants used to judge the similarity of the actions. While we asked participants to report which criteria they used for categorization in a follow-up survey at the end of the third experimental session, an explorative review of their free-text responses did not provide a clear picture. The multiple regression analysis reported in Study III, in contrast, showed a consistent picture across studies. In both experiments, the M model (i.e., whether one or both hands are used) and the MD model (i.e., movement direction) were significant predictors for the MDS model. They explained 49% of the variance in the subjective similarity ratings in Experiment 1

and 50% in Experiment 2.[7] As defined, the MD model captured trajectory information. The M model, in turn, can be related to action complexity. An action that needs a second, stabilizing hand can be seen as more complex than a purely unimanual action, with coordination, interaction and the role of each hand being key aspects (Krebs & Asfour, 2022). In infants, mastering role-differentiated bimanual manipulations (e.g., object-directed actions in which one hand stabilizes and the other hand manipulates an object) is a developmental milestone showing interhemispheric coordination (Kimmerle et al., 2010). Hence, movement trajectory and action complexity seem to influence observers' similarity judgments to a huge degree. It is an interesting question for future research which factors further shape subjective action classification as the resulting model seems to capture similarity structures that are mirrored in large parts of the action observation network.

Our results can be related to previous studies that used different tasks, asked participants to rate objects, not actions, based on their manipulation similarity and found primary object information to drive the coupling between pMTG and aIPS while primary object knowledge seemed to be mostly grip-type related (Hussain et al., 2024). Similarly, Watson and Buxbaum (2014) found the configuration of the hand and the magnitude of arm movement to play a role in determining how objects (in this case tools) cluster in action semantic space. The latter could hint in a similar direction as discussed above; however, our arrangement task did not ask for object manipulation knowledge, instead, it asked for observed action similarity across objects. Accordingly, Tucciarelli et al. (2019) asked for action similarity in an inverse multidimensional scaling protocol to examine the representational space of observed actions and found that the LOTC best captured the semantic dissimilarity structure. The discrepancy with our results could be due to the different stimulus material used. Tucciarelli et al. (2019) used static images of mostly intransitive everyday actions, while we used videos of object-

[7] As a brief reminder, in Experiment 2, the TU model significantly explained another 5% of the variance and in experiment 1, the additionally explained variance of 2% did not reach significance.

directed actions. Eventually, decoding the representational space of observed actions based on subjective ratings invites huge differences based on material and instructions. Still, the fact that this approach can link subjective and neural representations of action knowledge promises great progress in the understanding of semantic representation in the brain. The big picture is still unfolding, with many outstanding questions yet to be answered and actively being addressed in the field.

4.4 The aIPL in Action

One brain region that we repeatedly encountered in the current work is the aIPL. It constitutes a component of the action observation network, so its involvement does not appear unexpectedly. Upon closer inspection, its functional profile across the present studies is diverse and that is why I dedicate a separate paragraph to it. First, it is important to note that I refer to the anterior part of the SMG and the ventral part of the postcentral sulcus (i.e., the aIPS) when saying aIPL. More specifically, according to the Human Connectome Project (HCP) atlas as presented in Rolls et al. (2023), I refer to region PFt (and maybe anterior PF). The aIPL ROI used in Study II was created based on area PFt (Caspers et al., 2006, 2008) of the Julich-Brain Cytoarchitectonic Atlas (Amunts et al., 2020; Eickhoff et al., 2005). All reported whole-brain activation clusters in aIPS and aSMG of Study I and II peak within this ROI, and also the center coordinates of the spherical aIPS ROIs in Study III fall into this region.

To briefly recall the results per hemisphere, we found objective action categories to be represented in the left aIPL and the left aIPL was found active at subjective boundaries for dough items. We found the right aIPL to be active at subjective boundaries for commonplace items, at touchings for dough versus commonplace items, and to represent the subjective action space. Furthermore, we found bilateral aIPL activity to be increased for commonplace versus dough items during the entire length of the action videos, at untouchings for dough and dough versus commonplace items and finally to represent the manual model.

Possible hemispheric asymmetries or specializations of the aIPL are the topic of ongoing debates as many lateralized results suggest that but no definite conclusion could be made due to a multitude of factors (for reviews see e.g., Kemmerer, 2021; Tunik et al., 2007). For instance, when observing hand-object interactions, the identity of the acting hand, the visual hemifield where the action occurs, and the hand preference of the observer are just some of the many factors that need to be taken into account to interpret lateralized results. That said, please note that the participants in the reported studies were all right-handed and observed centrally presented actions performed by either the right or left hand (counterbalanced within and across participants).

Functionally, the aIPL has been described as a critical node within a network that dynamically controls actions on a higher order, that clearly exceeds low-level representations of grasp configurations and includes the representation of intended action goals (Tunik et al., 2007). Several studies suggest that the aIPL derives knowledge based on the identity of objects (Bach et al., 2010; Grafton & Hamilton, 2007; Urgen & Orban, 2021). Regarding the two main visual pathways, the aIPL belongs to the dorsal visual stream and, more precisely, it has been suggested to belong to the ventro-dorsal stream that is concerned with knowledge about object-associated actions (Binkofski & Buxbaum, 2013). Relatedly, Liu et al. (2024) found the bilateral aIPL (and right ventral premotor cortex) to encode action goals independent of action outcomes (i.e., independent of whether the action succeeds in reaching the desired end-state, such as an open bottle when opening a bottle) at an object-specific level. At the same time, only the left aIPL also contained goal information at an object-independent level.

The prominent role of object identities that emerges prompts me to review the univariate results for commonplace and dough items separately, despite possible redundancy with previous sections. For commonplace items that come with rich object-related knowledge, the aIPL activity was increased compared to dough items at the global video level and increased

aIPL activity had been yielded at subjective boundaries, though the latter did not survive the comparison to subjective boundary processing in dough videos. The global activity increase is in line with previous results indicating increasing aIPL activity for an increasing number of object-related actions (Schubotz et al., 2014) and was hypothesized in Study II. For dough items that come with limited object-related associations, aIPL effects were found at all three event boundary types: on whole-brain level at subjective event boundaries (though, here again, this effect did not survive the comparison to the corresponding contrast in commonplace items), on ROI level at touchings versus non-boundaries (here, aIPL activity was less reduced in dough items than in commonplace items), and on whole-brain and ROI level at untouchings (both within dough items and compared to commonplace items). Regarding the effect at the point of untouching, extracted contrast estimates show that aIPL activity is increased (compared to non-boundaries) for dough items and, at the same time, decreased (compared to non-boundaries) for commonplace items, which amplifies the effect. As mentioned earlier, this pattern is interpreted as suggesting a restricted (or well-informed) candidate action space for familiar objects and an unrestricted candidate action space for objects of limited information value at the point where the next action step is predicted.

Regarding the multivariate results, the fact that the left aIPL was associated with the objective action space, the right aIPL was associated with the subjective action space and both were associated with the manual model, highlights the high level of discriminability of action classes at the parietal level of the action observation network, as previously shown by Urgen and Orban (2021). Consistent with Liu et al. (2024), the left aIPL was found to represent objective action categories that carry cross-object action information.

The key conclusions about the aIPL's role in action observation and event boundary processing are that experience-based object-related knowledge modulates its bilateral recruitment during action perception and prediction, representing objective action categories

across objects in the left hemisphere at the same time. Therefore, the aIPL is considered to process object-related actions based on action knowledge.

4.5 Critical Evaluation and Methodological Considerations

Despite the significant findings of the present studies, certain experimental aspects require critical evaluation. Some limitations and corresponding suggestions for improvement have already been discussed so that this section specifically focuses on the experimental paradigm and the stimulus material.

In designing the paradigm, it was important to us that the action videos were passively observed during functional scanning, reliably segmented in post-fMRI sessions and finally categorized, when the participants knew the video material very well. While the methodology employed is robust, some aspects could be improved. Implementing a behavioral test-retest procedure in action segmentation yielded reliable subjective event boundaries, though pressing a button during the ongoing presentation of a video introduces reaction time considerations. We did not subtract a hypothetical motor response as the participants knew the videos from the scanning session already. Furthermore, not the test session's but the retest session's responses determined the exact timing of the event boundaries, so the participants were familiar with the segmentation procedure and the videos. Hence, we adopted the premise that button presses were delivered in anticipation of critical events in the videos, not in a reactive manner. The idea was that anticipation generates an early onset of the response that is then cancelled out by the motor reaction delay so that the registered button press hits the intended time. It is evident that this premise invites discussion, even though we have ultimately used group-aggregated time points. To avoid this premise, it would have been advantageous to give participants control about the video playing. If they had had the ability to stop and rewind the video, they could have been very precise about when to mark a unit. Unfortunately, this comes with a significantly

higher time expenditure per action video and would not have been feasible with the large number of individual videos used in this work. In the reported experiments, individual action videos were not repeatedly presented with the aim of introducing natural variability in trajectories and timings. In future research, one could decide to reduce the number of individual videos, present them repeatedly during functional scanning and give participants more time per video to mark event boundaries. Although the unit marking procedure employed in the present work is a well-established method, we combined it with a novel approach which gives rise to new challenges. The precise timing of subjective event boundaries becomes increasingly critical when being related to objective event boundaries and being utilized in time-locked brain activity analyses. However, the latter is somewhat relativized by the coarse temporal resolution of fMRI. Future studies that leverage MEG or EEG to investigate objective and subjective event boundaries are recommended to pay particular attention to this aspect.

Another element of the paradigm that requires consideration is the categorization task in the last behavioral session. Since the participants knew the action videos very well at that point, the videos were represented by image triplets showing the start of the scene, the object manipulation in the middle of the video and the final position of the objects. However, it cannot be completely ruled out that presenting the videos (i.e., staying with the stimulus format) instead of showing image triplets would have been beneficial. Additionally, and more crucially, a bigger, responsive screen could have improved the setup. Some participants' difficulty in completing the multi-arrangement task within a 60-minute timeframe was partly due to the limited screen size which prevented presenting all stimuli simultaneously. Despite the 27-inch screen being considered large at the time of data collection, screen technology has evolved considerably since then. Currently, tabletops may incorporate large screens that are operable via touch input. Thus, participants could sit at a table and arrange the videos in the same way one would arrange playing cards, using their finger. The screen should be large enough to display all stimuli concurrently, allowing for their use in at least some trials. This will facilitate more efficient

comparisons and enhance the overall experience for the participants. The videos could be represented as image triplets, and touching the middle image would play the video. Embracing these suggestions leads to even better MDS models to be used in multivariate analyses or to be analyzed independently.

Finally, the stimuli could potentially be enhanced. First, regarding the commonplace items' manipulations of Experiment 1, the extent to which the objects suggested the applied action varied. As an example, adding the last piece of a wooden puzzle differs from turning a calculator. While the former is quite predictable from the start image where the last piece of the puzzle lies in front of the agent (put together action), the latter is less foreseeable even if the turning action gets the calculator in the right orientation to be used afterwards. An unpublished explorative post-study rating in a separate group of 10 participants confirmed this difference in expectability between object manipulations. This could be driven by the fact that completing a puzzle aligns with its functional goal while a calculator that is turned without being used to perform a calculation does not fulfill its primary function. Controlling this dimension would result in a meaningful improvement of our paradigm. On the other hand, systematic manipulation of this dimension could also be considered. This could be accomplished through ratings of how expected a specific action appears when an object is seen in context. Preliminary results of an unpublished exploratory comparison between object-implied actions and not-object-implied actions of Experiment 1 (binary coded based on the above-mentioned rating) are pointing in the same direction as the reported comparison of strong and weak object-action associations. To clarify, however, these are two different concepts. Whether the concrete action that is applied to an object is expected (maybe because it aligns with the functional goal of the object) differs from the strength of action associations that the object carries, though they are certainly not independent. The latter is inherent to the object (independent of the applied action) and was systematically examined in the current work.

The strength of object-action associations is the second aspect of the stimuli that future studies could adjust. While we employed a binary distinction between strong and weak associations, future work could investigate the effect of object-action association on a continuum. To this end, participants could list the actions they expect when seeing an object and rate the level of expectedness of each action. Based on the results of such a pre-study, stimulus material could be designed accordingly. Ideally, the objects to be rated should be presented exactly as they would be seen at the start of the action video (i.e., embedded in a scene) as the configuration and the context of a scene facilitate action recognition and render some actions less expected. This does not only refer to the location of the action but also the position of the hand to the object (peripersonal space) and the presence of other objects (El-Sourani et al., 2018, 2019; Kemmerer, 2021; Wurm & Schubotz, 2012).

Returning to the current stimulus material, in retrospect, it must be noted that in some cases the initial scene of a video predicted the unfolding action. Keeping the start scene constant across actions could help reduce the initial predictability. In the present experiments, the scenes differed at the beginning of the action videos. Even if there was no systematic covariance between the relative position of the object to the subject and the action category, the method could be improved by keeping the spatial position of the objects at the start of the scene constant. This would, however, imply more objects to be present in the scene, which has an effect on action perception (El-Sourani et al., 2018, 2019). At the same time, the variability in objects used in the first experiment should certainly be preserved as it ensures the independence of the actions from specific grip types. I would even recommend future studies to incorporate this dimension into the dough items. In Experiment 2, formed pieces of blue dough were manipulated to limit object-related associations. Yet, it could be beneficial to adopt the object variability. Each action could be performed with several differently formed pieces of dough, each requiring a different type of grip.

Finally, the employed stimuli have fulfilled their purpose to essentially demonstrate that objective event boundaries in the form of changing contact states between objects are valuable for understanding human action perception. They are, however, far from what we perceive in our everyday life. Naturalistic action perception is exponentially more multifaceted, which limits the scope of the present results. To build upon the current findings, future work could examine event structure perception using more complex actions. In the following section, I will outline key directions for future investigations.

4.6 Future Prospects

Derived from the discussed limitations new experimental approaches emerge. Several suggestions have already been outlined in the course of the discussion, so this section will focus on ecological validity and possible adjustments to the current paradigm for follow-up studies. Subsequently, I will briefly address open questions and the resulting research approaches.

Firstly, the ecological validity could be increased by altering action content and action presentation. Regarding the content, one could use action videos taken from everyday life like, for instance, a family breakfast or a board games scene. Those videos could then show longer and more complex (inter)actions. The more complex a scene, the more important and insightful it is to know where participants look at so that future paradigms can largely profit from using eye tracking. Especially in more complex action scenes, it is crucial to examine whether the attentional focus of the participants is drawn to the touchings and untouchings during action observation.

Increasing ecological validity comes with the advantage of increasing the relevance of real-world-based internal models. Adults are sensitive to context (see e.g., El-Sourani et al.,

2019; Kemmerer, 2021; Wurm et al., 2012; Wurm & Schubotz, 2017) and I assume that they were very fast in recognizing that the videos in the experimental session of the reported experiments were somehow artificial and that the experience-based models that they gathered in the real world might not be entirely suitable in this specific situation. Therefore, participants could have been generally more open to expect odd manipulations in the experimental context after having seen the first handful of stimuli. A context that resembles the real world so that real-world-based internal models are relevant can be beneficial to investigate the role of experience-based predictions in ongoing perception.

Regarding the presentation of the actions, VR technology offers an interesting opportunity. It allows to add naturalistic elements in varying degrees and to program and thus manipulate displayed movements as needed (for action observation in VR see e.g., Lakshminarayanan et al., 2023; Wörgötter et al., 2020; Ziaeetabar et al., 2020). In addition, participants could change their perspective on the scene (through head movement or even walking through the scene) and eye movements could directly be tracked, nonetheless. Moreover, in VR the presented action could be adaptive and react to observer behavior (e.g., stop when the gaze is averted). Recently, Pooja et al. (2024) offered some guidelines to design ecologically valid cognitive neuroscience studies of event cognition that offer interesting ideas for VR and augmented reality, and Schubotz et al. (2023) inspire forward-thinking methodological VR approaches for studying hand actions that involve the use of tools. Thus, VR is set to play a significant role in action observation research. However, it must be acknowledged that VR approaches are currently not MR-compatible (except to a very limited degree as in Adamovich et al., 2009) and other functional imaging techniques are also difficult to combine with them. Considering the rapid development in this field in the last decade, I am still confident that these obstacles can be resolved in the decade to come.

While VR is an exciting field, there are also valuable research opportunities that remain closely aligned to the original studies of this thesis. As mentioned earlier, to further investigate the value of objective event boundaries for subjective event perception, it is important to identify those TUs that trigger a subjective event boundary in differentiation to those that do not. While it is a huge advantage for computer vision systems to not have to identify objects, it might be advantageous to identify at least the effecting object (i.e., the hand or a tool). A follow-up study could dissociate the TUs that emerge between the hand and an arbitrary object from those TUs that emerge between objects and object and ground. For either class of TUs, the temporal co-occurrence with subjective event boundaries could be examined. In the next phase, it may be exciting to replace hands and objects with animated cubes to help understand whether it is the hand (as part of the human body) or its functional role as an effector that assigns a special status to it (if this was the case). Regarding the underlying neural processing of TU incidents, time-locked analyses could be dissociated for hand-object-TUs to explore whether brain responses differ between the two classes of TUs. Returning to the computer-vision perspective, it could be worth testing whether it improves the algorithm to identify the effector in an action (e.g., by labelling the one that moves independently).

In addition, it is worth deepening the understanding of the effect of object-related knowledge. For future research, it can be a useful addition to manipulate object information on a continuum and to set up a rating study measuring also other concepts like affordance, functional knowledge, object familiarity, and object complexity. As previously done, the number of actions associated with the object could be assessed (Schubotz et al., 2014) along with whether the action performed on the object in the concrete stimulus is expected. This would enable us to gain a differentiated insight into the effect of object-action associations and their correlation to other concepts.

Finally, the neural activation patterns underlying event boundary processing need further empirical work to clarify the functional role of individual regions. Especially the aIPL merits additional attention. In this context, more diverse stimulus material could enable future studies to disentangle which brain areas are generally involved in event structure perception and which areas are content-specific, also elucidating their interaction. Equally important, the prominent role of untouchings, compared to touchings and observer-labeled event boundaries, deserves in-depth investigation. Considering more extended and complex actions, it is exciting to see whether untouchings continue to serve as the temporal anchor for prediction. Regarding the representation of subjectively derived action spaces, the current work merely pointed toward an initial understanding. The multitude of brain regions that are significantly associated with it warrants a closer examination, however. Particularly the dimensions underlying the subjective action space could be investigated in greater depth. To this end, future research can draw inspiration from studies of object recognition that create similarity spaces.

Looking ahead, the most promising methodological aspects for future research on event structure perception in action observation lie in increasing ecological validity, identifying key modulating factors and taking advantage of multidisciplinary approaches. In terms of topical aspects, future investigations should aim to further ground subjective event annotations in objective stimulus features, examine domain-general boundary-evoked activation patterns, the underlying dimensions of action space, and their related representation.

5 Conclusion

This thesis aimed to examine event perception, being a central cognitive mechanism, structuring time in a way that enables the brain to form memories and generate predictions. It investigated whether objective stimulus features in the form of touchings and untouchings between objects, hands and the ground drive subjective event segmentation and help understanding event structure perception. Furthermore, the effect of object-associated knowledge on event structure perception in object-directed actions was considered. At the same time, the value of objective action categories for understanding neural action representation was assessed and subjective action categories explored.

The findings indicate that objective event boundaries are a meaningful addition to subjective event boundaries to understand event structure perception in object-directed actions. They offer objective anchor points for behavioral action segmentation and help disentangling the neural signatures of event structure. Specifically, subjectively annotated event boundaries were mainly motion-driven and at the point of touching, low-level visual inspection of the scene intensified. The moment when objects un-touch proved to be crucial for attentional recalibration, memory encoding and predicting the upcoming action step. This prominent role of untouchings renders them important objective event boundaries in the context of predictive action processing.

Event structure processing was influenced by the wealth of information an object provided. Limited object associations rendered subjective boundaries even closer to objective boundaries and movement information weighed heavier when predicting the upcoming action at untouching. Simultaneously, limited object associations led to an unrestricted search for candidate actions. Conversely, rich object associations continuously activated associated actions and the rich visual information offered by commonplace items dominated processing at action step synchronization.

Finally, objective as well as subjective action categories were represented in brain regions belonging to the action observation network. The subjective action space was associated with a broad bilateral network while objective action categories were associated to the representational profile of a single brain area, that is, the aIPL. This brain area became a key region of the current work, and it was elaborated upon the significance of its prominent role in object-associated action knowledge and action class processing. The results suggest that the aIPL is involved in predicting object-related actions based on associated action knowledge.

This thesis contributes to the literature by offering a new perspective on event structure perception, combining computer vision with cognitive neuroscience. The inclusion of objective, stimulus-derived event boundaries allowed a structured view on the neural processes underlying ongoing event perception. The prominent role of objective boundaries in predictive processes underscores their fundamental purpose. Moreover, objective action categories revealed important insights regarding the functional role of the anterior intraparietal sulcus. In sum, this research offers valuable insights, although the scope is limited due to the decontextualized action stimuli.

The current results suggest important implications for human-robot cooperation as they could allow autonomous systems to make reliable predictions about human action. They could be of practical value for real-world applications in commercial robotics, such as home assistance technologies. More importantly, they can be relevant for clinical applications. They could help optimize training protocols used to restore function in neurorehabilitation and inform the development of robotic systems designed to support or train patients with motor impairments.

Future studies should identify the key objective events and investigate these in real-world-like scenarios, inviting interdisciplinary cooperation. Furthermore, boundary-evoked activation patterns warrant further attention as well as the effects of different contextual

aspects. Finally, additional research is required to explore the factors that shape categorical knowledge spaces.

Overall, this thesis provides a foundational understanding of objective, stimulus-derived features that drive event structure perception and corresponding categories' representation, taking the modulation by experience-based knowledge into account. It paves the way for more detailed investigations into neural event structure processing to eventually comprehend this central ability that essentially organizes experience.

6 References

Adamovich, S. V., August, K., Merians, A., & Tunik, E. (2009). A virtual reality-based system integrated with fmri to study neural mechanisms of action observation-execution: A proof of concept study. *Restorative Neurology and Neuroscience*, *27*(3), 209–223. https://doi.org/10.3233/RNN-2009-0471

Ahlheim, C., Stadler, W., & Schubotz, R. I. (2014). Dissociating dynamic probability and predictability in observed action - an fMRI study. *Frontiers in Human Neuroscience*, *8*. https://doi.org/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. *International Journal of Robotics Research*, *30*(10), 1229–1249. https://doi.org/10.1177/0278364911410459

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*(8), 379–390. https://doi.org/10.1016/j.tics.2013.06.009

Amunts, K., Mohlberg, H., Bludau, S., & Zilles, K. (2020). Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science*, *369*(6506), 988–992. https://doi.org/10.1126/science.abb4588

Bach, P., Peelen, M. V., & Tipper, S. P. (2010). On the role of object information in action observation: An fMRI study. *Cerebral Cortex*, *20*(12), 2798–2809. https://doi.org/10.1093/cercor/bhq026

Bailey, H. R., Kurby, C. A., Sargent, J. Q., & Zacks, J. M. (2017). Attentional focus affects how events are segmented and updated in narrative reading. *Memory and Cognition*, *45*(6), 940–955. https://doi.org/10.3758/s13421-017-0707-2

Bailey, H., & Smith, M. E. (2024). Event perception and event memory in real-world experience. In *Nature Reviews Psychology*. Nature Publishing Group. https://doi.org/10.1038/s44159-024-00367-0

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407. https://doi.org/10.1016/j.cognition.2007.07.005

Barnett, A. J., Nguyen, M., Spargo, J., Yadav, R., Cohn-Sheehy, B. I., & Ranganath, C. (2024). Hippocampal-cortical interactions during event boundaries support retention of complex narrative events. *Neuron*, *112*(2), 319-330.e7. https://doi.org/10.1016/j.neuron.2023.10.010

Ben-Yakov, A., & Henson, R. N. (2018). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience*, *38*(47), 10057–10068. https://doi.org/10.1523/JNEUROSCI.0524-18.2018

Betti, V., DellaPenna, S., de Pasquale, F., Mantini, D., Marzetti, L., Romani, G. L., & Corbetta, M. (2013). Natural scenes viewing alters the dynamics of functional connectivity in the human brain. *Neuron*, *79*(4), 782–797. https://doi.org/10.1016/j.neuron.2013.06.022

Biagi, L., Cioni, G., Fogassi, L., Guzzetta, A., Sgandurra, G., & Tosetti, M. (2016). Action observation network in childhood: a comparative fMRI study with adults. *Developmental Science*, *19*(6), 1075–1086. https://doi.org/10.1111/desc.12353

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Binkofski, F., & Buxbaum, L. J. (2013). Two action systems in the human brain. *Brain and Language*, *127*(2), 222–229. https://doi.org/10.1016/j.bandl.2012.07.007

Bledowski, C., Rahm, B., & Rowe, J. B. (2009). What "works" in working memory? Separate systems for selection and updating of critical information. *Journal of Neuroscience*, *29*(43), 13735–13741. https://doi.org/10.1523/JNEUROSCI.2547-09.2009

Canel, C., Kim, T., Zhou, G., Li, C., Lim, H., Andersen, D. G., Kaminsky, M., & Dulloor, S. R. (2019). Scaling Video Analytics on Constrained Edge Nodes. *Proceedings of Machine Learning and Systems 1*, 406–417.

Caspers, S., Eickhoff, S. B., Geyer, S., Scheperjans, F., Mohlberg, H., Zilles, K., & Amunts, K. (2008). The human inferior parietal lobule in stereotaxic space. *Brain Structure and Function*, *212*(6), 481–495. https://doi.org/10.1007/s00429-008-0195-z

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., & Zilles, K. (2006). The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. *NeuroImage*, *33*(2), 430–448. https://doi.org/10.1016/j.neuroimage.2006.06.054

Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage*, *50*(3), 1148–1167. https://doi.org/10.1016/j.neuroimage.2009.12.112

Cerliani, L., Bhandari, R., De Angelis, L., van der Zwaag, W., Bazin, P. L., Gazzola, V., & Keysers, C. (2022). Predictive coding during action observation – A depth-resolved intersubject functional correlation study at 7T. *Cortex*, *148*, 121–138. https://doi.org/10.1016/j.cortex.2021.12.008

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2024). Hacking the Predictive Mind †. *Entropy*, *26*(8). https://doi.org/10.3390/e26080677

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. https://doi.org/10.1038/nn.3381

Cutting, J. E. (2005). Perceiving scenes in film and in the world. *Moving Image Theory: Ecological Considerations*, 9–27.

De Soares, A., Kim, T., Mugisho, F., Zhu, E., Lin, A., Zheng, C., & Baldassano, C. (2024). Top-down attention shifts behavioral and neural event boundaries in narratives with overlapping event scripts. *Current Biology*, *34*(20), 4729-4742.e5. https://doi.org/10.1016/j.cub.2024.09.013

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Experimental Brain Research Understanding motor events: a neurophysiological study. In *Exp Brain Res* (Vol.

91).

Dima, D. C., Janarthanan, S., Culham, J. C., & Mohsenzadeh, Y. (2024). Shared representations of human actions across vision and language. *Neuropsychologia*, *202*. https://doi.org/10.1016/j.neuropsychologia.2024.108962

DuBrow, S. (2024). Events and Boundaries. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory* (pp. 497–519). Oxford University Press Inc.

DuBrow, S., & Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory*, *134*, 107–114. https://doi.org/10.1016/j.nlm.2016.07.011

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0146-z

Eiteljoerge, S. F. V., Adam, M., Elsner, B., & Mani, N. (2019). Word-object and action-object association learning across early development. *PLoS ONE*, *14*(8), 1–22. https://doi.org/10.1371/journal.pone.0220317

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. *Journal of Cognitive Neuroscience*, *32*(2), 326–337. https://doi.org/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*(November 2017), 429–437. https://doi.org/10.1016/j.neuroimage.2017.11.047

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, *22*(2), 243–252. https://doi.org/10.1177/0956797610393742

Ezzyat, Y., & Davachi, L. (2021). Neural Evidence for Representational Persistence Within Events. *The Journal of Neuroscience*, *41*(37), 7909–7920. https://doi.org/10.1523/JNEUROSCI.0073-21.2021

Faber, M., Radvansky, G. A., & D'Mello, S. K. (2018). Driven to distraction: A lack of change gives rise to mind wandering. *Cognition*, *173*, 133–137. https://doi.org/10.1016/j.cognition.2018.01.007

Falck-Ytter, T., Gredebäck, G., & Von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878–879. https://doi.org/10.1038/nn1729

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

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771. https://doi.org/10.1038/nature01861

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal Lobe:

From Action Organization to Intention Understanding. *Science*, *308*(5722), 662–667.

Fox, R., & McDaniel, C. (1982). The perception of biological motion by human infants. *Science*, *218*(4571), 486–487.

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). *Structured event memory: a neuro-symbolic model of event cognition*. https://doi.org/10.1101/541607

Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A., & Rogers, T. T. (2023). Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences*, *27*(3), 258–281. https://doi.org/10.1016/j.tics.2022.12.006

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

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

Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex*, *19*(6), 1239–1255. https://doi.org/10.1093/cercor/bhn181

Gesierich, B., Bruzzo, A., Ottoboni, G., & Finos, L. (2008). Human gaze behaviour during action execution and observation. *Acta Psychologica*, *128*(2), 324–330. https://doi.org/10.1016/j.actpsy.2008.03.006

Geva-Sagiv, M., Dimsdale-Zucker, H. R., Williams, A. B., & Ranganath, C. (2023). Proximity to boundaries reveals spatial context representation in human hippocampal CA1. *Neuropsychologia*, *189*. https://doi.org/10.1016/j.neuropsychologia.2023.108656

Grafton, S. T., & Hamilton, A. F. D. C. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, *26*(4), 590–616. https://doi.org/10.1016/j.humov.2007.05.009

Hahamy, A., Dubossarsky, H., & Behrens, T. E. J. (2023). The human brain reactivates context-specific past information at event boundaries of naturalistic experiences. *Nature Neuroscience*, *26*(6), 1080–1089. https://doi.org/10.1038/s41593-023-01331-6

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586–604. https://doi.org/10.1037/a0024310

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory and Cognition*, *34*(6), 1221–1235. https://doi.org/10.3758/BF03193267

Hardwick, R. M., Caspers, S., Eickhoff, S. B., & Swinnen, S. P. (2018). Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience & Biobehavioral Reviews*, *94*, 31–44. https://doi.org/10.1016/j.neubiorev.2018.08.003

Hodgson, V. J., Lambon Ralph, M. A., & Jackson, R. L. (2023). The cross-domain functional organization of posterior lateral temporal cortex: insights from ALE meta-analyses of 7 cognitive domains spanning 12,000 participants. *Cerebral Cortex*, *33*(8), 4990–5006. https://doi.org/10.1093/cercor/bhac394

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

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), 1–12. https://doi.org/10.1371/journal.pone.0134316

Huff, M., Meitz, T. G. K., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(5), 1377–1388. https://doi.org/10.1037/a0036780

Hunnius, S., & Bekkering, H. (2010). The Early Development of Object Knowledge: A Study of Infants' Visual Anticipations During Action Observation. *Developmental Psychology*, *46*(2), 446–454. https://doi.org/10.1037/a0016543

Hunnius, S., & Bekkering, H. (2014). What are you doing? How active and observational experience shape infants' action understanding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1644), 20130490. https://doi.org/10.1098/rstb.2013.0490

Hussain, A., Walbrin, J., Tochadse, M., & Almeida, J. (2024). Primary manipulation knowledge of objects is associated with the functional coupling of pMTG and aIPS. *Neuropsychologia*, *205*. https://doi.org/10.1016/j.neuropsychologia.2024.109034

Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L., & Santamaría, J. (2022). Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics*, *12*(1), 29. https://doi.org/10.3390/electronics12010029

Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action Genome: Actions as Compositions of Spatio-temporal Scene Graphs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10236–10247.

Kalénine, S., Wamain, Y., Decroix, J., & Coello, Y. (2016). Conflict between object structural and functional affordances in peripersonal space. *Cognition*, *155*, 1–7. https://doi.org/10.1016/j.cognition.2016.06.006

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2016). The occipital place area represents the local elements of scenes. *NeuroImage*, *132*, 417–424. https://doi.org/10.1016/j.neuroimage.2016.02.062

Kaplan, J. T., & Meyer, K. (2012). Multivariate pattern analysis reveals common neural patterns across individuals during touch observation. *NeuroImage*, *60*(1), 204–212. https://doi.org/10.1016/j.neuroimage.2011.12.059

Kemmerer, D. (2021). What modulates the Mirror Neuron System during action observation? *Progress in Neurobiology*, *205*, 102128. https://doi.org/10.1016/j.pneurobio.2021.102128

Keysers, C., Silani, G., & Gazzola, V. (2024). Predictive coding for the actions and emotions of others and its deficits in autism spectrum disorders. *Neuroscience & Biobehavioral Reviews*, *167*, 105877. https://doi.org/10.1016/j.neubiorev.2024.105877

Kilner, J. M. (2011). More than one pathway to action understanding. In *Trends in Cognitive Sciences* (Vol. 15, Issue 8, pp. 352–357). https://doi.org/10.1016/j.tics.2011.06.005

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. In *Cognitive Processing* (Vol. 8, Issue 3, pp. 159–166). https://doi.org/10.1007/s10339-007-0170-2

Kimmerle, M., Ferre, C. L., Kotwica, K. A., & Michel, G. F. (2010). Development of role-differentiated bimanual manipulation during the infant's first year. *Developmental Psychobiology*, *52*(2), 168–180. https://doi.org/10.1002/dev.20428

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Kong, Y., & Fu, Y. (2022). Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, *130*(5), 1366–1401. https://doi.org/10.1007/s11263-022-01594-9

Kosie, J. E., & Baldwin, D. (2019). Attentional profiles linked to event segmentation are robust to missing information. *Cognitive Research: Principles and Implications*, *4*(1). https://doi.org/10.1186/s41235-019-0157-4

Krebs, F., & Asfour, T. (2022). A Bimanual Manipulation Taxonomy. *IEEE Robotics and Automation Letters*, *7*(4), 11031–11038. https://doi.org/10.1109/LRA.2022.3196158

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. *Frontiers in Psychology*, *3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00245

Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2023). Bayesian Surprise Predicts Human Event Segmentation in Story Listening. *Cognitive Science*, *47*(10). https://doi.org/10.1111/cogs.13343

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. https://doi.org/10.1016/j.tics.2007.11.004

Lakshminarayanan, K., Shah, R., Daulat, S. R., Moodley, V., Yao, Y., & Madathil, D. (2023). The effect of combining action observation in virtual reality with kinesthetic motor imagery on cortical activity. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1201865

Leshinskaya, A., & Caramazza, A. (2015). Abstract categories of functions in anterior parietal lobe. *Neuropsychologia*, *76*, 27–40. https://doi.org/10.1016/j.neuropsychologia.2015.01.014

Lesourd, M., Afyouni, A., Geringswald, F., Cignetti, F., Raoul, L., Sein, J., Nazarian, B., Anton, J. L., & Grosbras, M. H. (2023). Action Observation Network Activity Related to Object-Directed and Socially-Directed Actions in Adolescents. *Journal of Neuroscience*, *43*(1), 125–141. https://doi.org/10.1523/JNEUROSCI.1602-20.2022

Levine, D., Buchsbaum, D., Hirsh-Pasek, K., & Golinkoff, R. M. (2019). Finding events in a continuous world: A developmental account. *Developmental Psychobiology*, *61*(3), 376–389. https://doi.org/10.1002/dev.21804

Li, M., Lu, S., & Zhong, N. (2016). The parahippocampal cortex mediates contextual associative memory: Evidence from an fMRI study. *BioMed Research International*, *2016*. https://doi.org/10.1155/2016/9860604

Lingnau, A., & Downing, P. (2024). Action Understanding. In *Action Understanding*. Cambridge

University Press. https://doi.org/10.1017/9781009386630

Liu, S., Wurm, M. F., & Caramazza, A. (2024). Dissociating goal from outcome during action observation. *Cerebral Cortex*, *34*(12). https://doi.org/10.1093/cercor/bhae487

Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent. *Developmental Science*, *14*(2), 453–460. https://doi.org/10.1111/j.1467-7687.2010.00995.x

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, *35*(8), 1489–1517. https://doi.org/10.1111/j.1551-6709.2011.01202.x

Malone, P. S., Glezer, L. S., Kim, J., Jiang, X., & Riesenhuber, M. (2016). Multivariate pattern analysis reveals category-related organization of semantic representations in anterior temporal cortex. *Journal of Neuroscience*, *36*(39), 10089–10096. https://doi.org/10.1523/JNEUROSCI.1599-16.2016

Meyer, K., Kaplan, J. T., Essex, R., Damasio, H., & Damasio, A. (2011). Seeing touch is correlated with content-specific activity in primary somatosensory cortex. *Cerebral Cortex*, *21*(9), 2113–2121. https://doi.org/10.1093/cercor/bhq289

Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2025). Large language models can segment narrative events similarly to humans. *Behavior Research Methods*, *57*(1), 39. https://doi.org/10.3758/s13428-024-02569-z

Morales, S., Bowman, L. C., Velnoskey, K. R., Fox, N. A., & Redcay, E. (2019). An fMRI study of action observation and action execution in childhood. *Developmental Cognitive Neuroscience*, *37*. https://doi.org/10.1016/j.dcn.2019.100655

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38. https://doi.org/10.1037/h0035584

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436–450. https://doi.org/10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847–862. https://doi.org/10.1037/0022-3514.35.12.847

Ogg, M., & Slevc, L. R. (2019). Acoustic Correlates of Auditory Object and Event Perception: Speakers, Musical Timbres, and Environmental Sounds. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01594

Palejwala, A. H., Dadario, N. B., Young, I. M., O'Connor, K., Briggs, R. G., Conner, A. K., O'Donoghue, D. L., & Sughrue, M. E. (2021). Anatomy and White Matter Connections of the Lingual Gyrus and Cuneus. *World Neurosurgery*, *151*, e426–e437. https://doi.org/10.1016/j.wneu.2021.04.050

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Pettijohn, K. A., & Radvansky, G. A. (2016). Narrative event boundaries, reading times, and expectation. *Memory and Cognition*, *44*(7), 1064–1075. https://doi.org/10.3758/s13421-016-0619-6

Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., & Radvansky, G. A. (2016). Event boundaries and memory improvement. *Cognition*, *148*, 136–144. https://doi.org/10.1016/j.cognition.2015.12.013

Pomp, J., Garlichs, A., Kulvicius, T., Tamosiunaite, M., Wurm, M. F., Zahedi, A., Wörgötter, F., & Schubotz, R. I. (2024). Action Segmentation in the Brain: The Role of Object–Action Associations. *Journal of Cognitive Neuroscience*, *36*(9), 1784–1806. https://doi.org/10.1162/jocn_a_02210

Pomp, J., Heins, N., Trempler, I., Kulvicius, T., Tamosiunaite, M., Mecklenbrauck, F., Wurm, M. F., Wörgötter, F., & Schubotz, R. I. (2021). Touching events predict human action segmentation in brain and behavior. *NeuroImage*, *243*, 118534. https://doi.org/10.1016/j.neuroimage.2021.118534

Pomp, J., Wurm, M. F., Selvan, R. N., Wörgötter, F., & Schubotz, R. I. (2025). Touching-untouching patterns organize action representation in the inferior parietal cortex. *NeuroImage*, *310*. https://doi.org/10.1016/j.neuroimage.2025.121113

Pooja, R., Ghosh, P., & Sreekumar, V. (2024). Towards an ecologically valid naturalistic cognitive neuroscience of memory and event cognition. *Neuropsychologia*, *203*. https://doi.org/10.1016/j.neuropsychologia.2024.108970

Pradhan, R., & Kumar, D. (2022). Event segmentation and event boundary advantage: Role of attention and postencoding processing. *Journal of Experimental Psychology: General*, *151*(7), 1542–1555. https://doi.org/10.1037/xge0001155

Radvansky, G. A., & Zacks, J. M. (2011). Event perception. *WIREs Cognitive Science*, *2*(6), 608–620. https://doi.org/10.1002/wcs.133

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. https://doi.org/10.1016/j.cobeha.2017.08.006

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2(1)*, 79–87.

Reagh, Z. M., Delarazan, A. I., Garber, A., & Ranganath, C. (2020). Aging alters neural activity at event boundaries in the hippocampus and Posterior Medial network. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17713-4

Reilly, J., Shain, C., Borghesani, V., Kuhnke, P., Vigliocco, G., Peelle, J. E., Mahon, B. Z., Buxbaum, L. J., Majid, A., Brysbaert, M., Borghi, A. M., De Deyne, S., Dove, G., Papeo, L., Pexman, P. M., Poeppel, D., Lupyan, G., Boggio, P., Hickok, G., … Vinson, D. (2025). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review*, *32*(1), 243–280. https://doi.org/10.3758/s13423-024-02556-7

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–643. https://doi.org/10.1080/15326900701399913

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141.

Rolls, E. T., Deco, G., Huang, C. C., & Feng, J. (2023). The human posterior parietal cortex: effective connectome, and its relation to function. *Cerebral Cortex*, *33*(6), 3142–3170.

https://doi.org/10.1093/cercor/bhac266

Ruotolo, F., Ruggiero, G., Raemaekers, M., Iachini, T., van der Ham, I. J. M., Fracasso, A., & Postma, A. (2019). Neural correlates of egocentric and allocentric frames of reference combined with metric and non-metric spatial relations. *Neuroscience*, *409*, 235–252. https://doi.org/10.1016/j.neuroscience.2019.04.021

Sacheli, L. M., Verga, C., Zapparoli, L., Seghezzi, S., Tomasetig, G., Banfi, G., & Paulesu, E. (2023). When action prediction grows old: An fMRI study. *Human Brain Mapping*, *44*(2), 373–387. https://doi.org/10.1002/hbm.26049

Sargent, J. Q., Zacks, J. M., & Bailey, H. R. (2015). Perceptual segmentation of natural events: Theory, methods, and applications. *The Cambridge Handbook of Applied Perception Research*, 443–465. https://doi.org/10.1017/CBO9780511973017.029

Sasmita, K., & Swallow, K. M. (2023). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*, *55*(1), 428–447. https://doi.org/10.3758/s13428-022-01832-5

Schiffer, A. M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (2012). Surprised at all the entropy: Hippocampal, caudate and midbrain contributions to learning from prediction errors. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0036445

Schubotz, R. I. (2015). Prediction and Expectation. In *Brain Mapping* (Vol. 3, pp. 295–302). Elsevier. https://doi.org/10.1016/B978-0-12-397025-1.00247-5

Schubotz, R. I., Ebel, S. J., Elsner, B., Weiss, P. H., & Wörgötter, F. (2023). Tool mastering today – an interdisciplinary perspective. In *Frontiers in Psychology* (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fpsyg.2023.1191792

Schubotz, R. I., Korb, F. M., Schiffer, 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. https://doi.org/10.1016/j.neuroimage.2012.04.008

Schubotz, R. I., & von Cramon, D. Y. (2001). Functional organization of the lateral premotor cortex: fMRI reveals different regions activated by anticipation of object properties, location and speed. *Cognitive Brain Research*, *11*(1), 97–112. https://doi.org/10.1016/S0926-6410(00)00069-0

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. *Frontiers in Psychology*, *5*, 1–15. https://doi.org/10.3389/fpsyg.2014.00636

Schulz, L., Ischebeck, A., Wriessnegger, S. C., Steyrl, D., & Müller-Putz, G. R. (2018). Action affordances and visuo-spatial complexity in motor imagery: An fMRI study. *Brain and Cognition*, *124*, 37–46. https://doi.org/10.1016/j.bandc.2018.03.012

Selvan, R. N., Cheng, M., Siestrup, S., Mecklenbrauck, F., Jainta, B., Pomp, J., Zahedi, A., Tamosiunaite, M., Wörgötter, F., & Schubotz, R. I. (2024). Updating predictions in a complex repertoire of actions and its neural representation. *NeuroImage*, *296*. https://doi.org/10.1016/j.neuroimage.2024.120687

Shin, Y. S., & DuBrow, S. (2021). Structuring Memory Through Inference-Based Event Segmentation. *Topics in Cognitive Science*, *13*(1), 106–127. https://doi.org/10.1111/tops.12505

Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective and Behavioral Neuroscience*, *3*(4), 335–345. https://doi.org/10.3758/CABN.3.4.335

Stadler, M. A., & Frensch, P. A. (1998). *Handbook of implicit learning*. Sage Publications, Inc.

Stadler, W., Ott, D. V. M., Springer, A., Schubotz, R. I., Schütz-Bosbach, S., & Prinz, W. (2012). Repetitive TMS suggests a role of the human dorsal premotor cortex in action prediction. *Frontiers in Human Neuroscience*, *FEBRUARY 2012*. https://doi.org/10.3389/fnhum.2012.00020

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. *Human Brain Mapping*, *32*(5), 677–687. https://doi.org/10.1002/hbm.20949

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

Swallow, K. M., & Zacks, J. M. (2008). Sequences learned without awareness can orient attention during the perception of human activity. *Psychonomic Bulletin and Review*, *15*(1), 116–122. https://doi.org/10.3758/PBR.15.1.116

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event Boundaries in Perception Affect Memory Encoding and Updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257. https://doi.org/10.1037/a0015631

Tucciarelli, R., Wurm, M., Baccolo, E., & Lingnau, A. (2019). The representational space of observed actions. *ELife*, *8*. https://doi.org/10.7554/eLife.47686

Tunik, E., Rice, N. J., Hamilton, A., & Grafton, S. T. (2007). Beyond grasping: Representation of action in human anterior intraparietal sulcus. *NeuroImage*, *36*, T77–T86. https://doi.org/10.1016/j.neuroimage.2007.03.026

Tversky, B., Zacks, J. M., & Lee, P. (2004). Events by hands and feet. *Spatial Cognition and Computation*, *4*(1), 5–14. https://doi.org/10.1207/s15427633scc0401_2

Urgen, B. A., & Orban, G. A. (2021). The unique role of parietal cortex in action observation: Functional organization for communicative and manipulative actions. *NeuroImage*, *237*. https://doi.org/10.1016/j.neuroimage.2021.118220

Urgen, B. A., Pehlivan, S., & Saygin, A. P. (2016). Representational similarity of actions in the human brain. *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1–4.

Urgen, B. A., & Saygin, A. P. (2020). Predictive processing account of action perception: Evidence from effective connectivity in the action observation network. *Cortex*, *128*, 132–142. https://doi.org/10.1016/j.cortex.2020.03.014

Watson, C. E., & Buxbaum, L. J. (2014). Uncovering the architecture of action semantics. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1832–1848. https://doi.org/10.1037/a0037449

Williams, J. N. (2020). The Neuroscience of Implicit Learning. *Language Learning*, *70*, 255–307. https://doi.org/10.1111/lang.12405

Worgotter, F., Aksoy, E. E., Kruger, N., Piater, J., Ude, A., & Tamosiunaite, M. (2013). A Simple Ontology of Manipulation Actions Based on Hand-Object Relations. *IEEE Transactions on Autonomous Mental Development*, *5*(2), 117–134.

https://doi.org/10.1109/TAMD.2012.2232291

Wörgötter, F., Ziaeetabar, F., Pfeiffer, S., Kaya, O., Kulvicius, T., & Tamosiunaite, M. (2020). Humans Predict Action using Grammar-like Structures. *Scientific Reports*, *10*(3999). https://doi.org/10.1038/s41598-020-60923-5

Wurm, M. F., Artemenko, C., Giuliani, D., & Schubotz, R. I. (2017). Action at its place: Contextual settings enhance action recognition in 4- to 8-year-old children. *Developmental Psychology*, *53*(4), 662–670. https://doi.org/10.1037/dev0000273

Wurm, M. F., & Caramazza, A. (2019). Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nature Communications*, *10*(1), 289. https://doi.org/10.1038/s41467-018-08084-y

Wurm, M. F., Caramazza, A., & Lingnau, A. (2017). Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *Journal of Neuroscience*, *37*(3), 562–575. https://doi.org/10.1523/JNEUROSCI.1717-16.2016

Wurm, M. F., Cramon, D. Y., & Schubotz, R. I. (2012). *The Context-Object-Manipulation Triad: Cross Talk during Action Perception Revealed by fMRI*. http://direct.mit.edu/jocn/article-pdf/24/7/1548/1944490/jocn_a_00232.pdf

Wurm, M. F., & Lingnau, A. (2015). Decoding Actions at Different Levels of Abstraction. *Journal of Neuroscience*, *35*(20), 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15. 2015

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

Wurm, M. F., & Schubotz, R. I. (2017). What's she doing in the kitchen? Context helps when actions are hard to recognize. *Psychonomic Bulletin and Review*, *24*(2), 503–509. https://doi.org/10.3758/s13423-016-1108-4

Wurm, M. F., & Schubotz, R. I. (2018). The role of the temporoparietal junction (TPJ) in action observation: Agent detection rather than visuospatial transformation. *NeuroImage*, *165*, 48–55. https://doi.org/10.1016/j.neuroimage.2017.09.064

Xiao, D., Dianati, M., Geiger, W. G., & Woodman, R. (2023). Review of Graph-Based Hazardous Event Detection Methods for Autonomous Driving Systems. *IEEE Transactions on Intelligent Transportation Systems*, *24*(5), 4697–4715. https://doi.org/10.1109/TITS.2023.3240104

Xu, C., Wang, J., Wan, K., Li, Y., & Duan, L. (2006). Live sports event detection based on broadcast video and web-casting text. *Proceedings of the 14th ACM International Conference on Multimedia*, 221–230. https://doi.org/10.1145/1180639.1180699

Xu, S., & Heinke, D. (2017). Implied between-object actions affect response selection without knowledge about object functionality. *Visual Cognition*, *25*(1–3), 152–168. https://doi.org/10.1080/13506285.2017.1330792

Xu, S., Humphreys, G. W., Mevorach, C., & Heinke, D. (2017). The involvement of the dorsal stream in processing implied actions between paired objects: A TMS study. *Neuropsychologia*, *95*, 240–249. https://doi.org/10.1016/j.neuropsychologia.2016.12.021

Yates, T. S., Sherman, B. E., & Yousif, S. R. (2023). More than a moment: What does it mean to call something an 'event'? *Psychonomic Bulletin & Review*, *30*(6), 2067–2082. https://doi.org/10.3758/s13423-023-02311-4

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2022). Neural event segmentation of continuous experience in human infants. *Proceedings of the National Academy of Sciences*, *119*(43). https://doi.org/10.1073/pnas.2200257119

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008. https://doi.org/10.1016/j.cogsci.2004.06.003

Zacks, J. M. (2020). Event Perception and Memory. *Annual Review of Psychology*, *71*(1), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human Brain Activity Time-Locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655. https://doi.org/https://doi.org/10.1038/88486

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216. https://doi.org/10.1016/j.cognition.2009.03.007

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction Error Associated with the Perceptual Segmentation of Naturalistic Events. *Journal of Cognitive Neuroscience*, *23*(12), 4057–4066. https://doi.org/10.1162/jocn_a_00078

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the alzheimer type. *Psychology and Aging*, *21*(3), 466–482. https://doi.org/10.1037/0882-7974.21.3.466

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84. https://doi.org/10.1111/j.1467-8721.2007.00480.x

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

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29–58. https://doi.org/10.1037/0096-3445.130.1.29

Zentgraf, K., Munzert, J., Bischoff, M., & Newman-Norlund, R. D. (2011). Simulation during observation of human actions – Theories, empirical studies, applications. *Vision Research*, *51*(8), 827–835. https://doi.org/10.1016/j.visres.2011.01.007

Zheng, Y., Zacks, J. M., & Markson, L. (2020). The development of event perception and memory. *Cognitive Development*, *54*. https://doi.org/10.1016/j.cogdev.2020.100848

Ziaeetabar, F., Pomp, J., Pfeiffer, S., El-Sourani, N., Schubotz, R. I., Tamosiunaite, M., & Wörgötter, F. (2020). Using enriched semantic event chains to model human action prediction based on (minimal) spatial information. *PLOS ONE*, *15*(12), e0243829. https://doi.org/10.1371/journal.pone.0243829

7 Abbreviations

AI	artificial intelligence
aIPL	anterior inferior parietal lobule
aIPS	anterior intraparietal sulcus
aSMG	anterior supramarginal gyrus
fMRI	functional magnetic resonance imaging
HCP	Human Connectome Project
LOC	lateral occipital cortex
LOTC	lateral occipitotemporal cortex
PHC	parahippocampal cortex
pMTG	posterior middle temporal gyrus
PMd	dorsal premotor cortex
ROI	region of interest
RSA	representational similarity analysis
rTMS	repetitive transcranial magnetic stimulation
SEC	semantic event chain
SMG	supramarginal gyrus
SPL	superior parietal lobule
T	touching incident (i.e., the moment when two objects touch)
TU	touching and untouching incidents
U	untouching incident (i.e., the moment when two objects un-touch)
VR	virtual reality