Behavior Research Methods (2023) 55:3629–3644
https://doi.org/10.3758/s13428-022-01980-8

The multi-angle extended three-dimensional activities (META) stimulus set: A tool for studying event cognition

Matthew A. Bezdek[1] · Tan T. Nguyen[1] · Christopher S. Hall[1] · Todd S. Braver[1] · Aaron F. Bobick[2] · Jeffrey M. Zacks[1]

Accepted: 12 September 2022 / Published online: 10 October 2022

Abstract

To study complex human activity and how it is perceived and remembered, it is valuable to have large-scale, well-characterized stimuli that are representative of such activity. We present the Multi-angle Extended Three-dimensional Activities (META) stimulus set, a structured and highly instrumented set of extended event sequences performed in naturalistic settings. Performances were captured with two color cameras and a Kinect v2 camera with color and depth sensors, allowing the extraction of three-dimensional skeletal joint positions. We tracked the positions and identities of objects for all chapters using a mixture of manual coding and an automated tracking pipeline, and hand-annotated the timings of high-level actions. We also performed an online experiment to collect normative event boundaries for all chapters at a coarse and fine grain of segmentation, which allowed us to quantify event durations and agreement across participants. We share these materials publicly to advance new discoveries in the study of complex naturalistic activity.

Keywords Action perception · Event cognition · Event segmentation · Naturalistic stimuli · Norms

Continuous experiences are parcellated into events, both during ongoing comprehension and in memory. Investigating the mechanisms by which experiences are transformed into events is of interest for studying basic behavioral and neurocognitive psychological research (Butz et al., 2019; DuBrow & Davachi, 2016; Richmond & Zacks, 2017; Schapiro et al., 2013; Shin & DuBrow, 2021; Zacks, Braver, et al., 2001a; Zacks et al., 2007; Zacks & Swallow, 2007), early childhood development (Baldwin et al., 2001; Hespos et al., 2009; Levine et al., 2019; Saylor et al., 2007), healthy aging (Kurby & Zacks, 2018; Magliano et al., 2012; Sargent et al., 2013), and clinical populations (Eisenberg et al., 2016; Richmond et al., 2017; Sherrill & Magliano, 2017; Zacks et al., 2006; Zalla et al., 2004, 2013).

The study of events in perception, memory, action control, and reasoning is referred to as *event cognition* (Radvansky & Zacks, 2014). Event cognition research has seen vigorous activity in the last two decades. We believe this reflects the existence of many important phenomena in psychology and neuroscience that are hard to investigate fully without considering the dynamics and multimodal structure of events in experience. Research on *event segmentation* has established that events are segmented simultaneously by the mind and the brain on multiple timescales (Baldassano et al., 2017; Hasson et al., 2008; Zacks, Braver, et al., 2001a; Zacks, Tversky, et al., 2001b). This perceptual structure is important not only for understanding perception, but also for understanding memory. Memory, like perception, is organized into events, such that features belonging to a common event tend to be recalled as a unit, and such that relations between events are represented in memory (DuBrow & Davachi, 2016; Lichtenstein & Brewer, 1980; Rubin & Umanath, 2015). Event structure in perception determines event structure in memory: The event boundaries identified by viewers of an activity serve as units of organization in memory (Ezzyat & Davachi, 2011; Michelmann et al., 2021). These relationships lead to relationships between individual and group differences in perception and in memory: People who segment activity more effectively remember more (Sargent et al., 2013). They are also better able to perform everyday activities (Bailey et al., 2013).

Perception and memory are both guided by knowledge—knowledge about how categories of objects look and sound, about how people act and talk, about how things move and

✉ Matthew A. Bezdek
mbezdek@wustl.edu

[1] Department of Psychological and Brain Sciences, Washington University in St. Louis, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899, USA

[2] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA

change. Knowledge about events, in the form of *schemas* or *scripts,* is of notable importance for understanding how people make sense of their worlds (Abelson, 1981; Rumelhart, 1980). Adults depend heavily on event knowledge to act within and to remember everyday activities (Barbey et al., 2009; Graesser & Nakamura, 1982; Schacter et al., 2017). This knowledge develops during childhood (Fivush, 1997; Nelson, 1986), and it can be affected by neurological and psychiatric diseases (Grafman et al., 1993; Roll et al., 2017; Zacks & Sargent, 2010). Event knowledge guides the encoding of new activities by providing scaffolds for event encoding (Bonasia et al., 2018; Bransford & Johnson, 1972), and it guides memory for activity by organizing retrieval and by enabling inferences about missing information (Anderson & Pichert, 1978; Bower et al., 1979).

The investigation of event structure in perception and memory requires stimuli that reflect the sequential, dynamic structure of activity, the patterns of correlations across stimulus dimensions, and the category structure of events. Thus, a set of naturalistic extended event sequences is a potentially valuable resource for psychologists and neuroscientists. It may also be useful for addressing issues of interest to researchers in the fields of machine learning and artificial intelligence. In particular, this stimulus set provides naturalistic and structured data related to several research areas in computer vision, including: object detection, object recognition (Liang & Hu, 2015; Russakovsky et al., 2015), object interaction (Li et al., 2019; Yao & Fei-Fei, 2010), scene segmentation (Fu et al., 2019; Zhang et al., 2018), action recognition (Kuehne et al., 2014; Simonyan & Zisserman, 2014), event recognition (Wang & Ji, 2015), and event segmentation (Aakur & Sarkar, 2019; Franklin et al., 2019).

Previously developed event corpora have made major contributions to machine vision (see Table 1). However, these resources lack some features that are valuable for psychological research and potentially for machine vision. Stimulus sets collected from the wild, such as YouTube videos (Monfort et al., 2019; Zhou et al., 2018) or Hollywood films (Kuehne et al., 2011; Marszalek et al., 2009), often provide higher realism at the expense of experimenter control.[1] Conversely, stimulus sets produced in a laboratory provide greater control and more precise measurements of the features of unfolding events, but often appear divorced from the realism of events as they are perceived in everyday life.

Of particular note, acquiring sequences of events that extend continuously for longer durations of time is challenging. For example, the Charades dataset (Sigurdsson et al., 2016) contains a large number of action videos collected through Amazon Mechanical Turk. This novel approach generated a large set of natural actions in diverse locations, which were exhaustively annotated. However, the brief durations of the action videos do not permit the study of continuous sequences of actions as they naturally follow each other. As shown in the average sequence length column in Table 1, the sequences captured in event stimulus sets tend to be brief.

To address these limitations, we created the Multi-angle Extended Three-dimensional Activities (META) stimulus set. Our aim was to create materials that would be useful for a broad range of applications in fields that study dynamic complex event sequences, including psychology and cognitive neuroscience. We sought to create highly realistic, continuous event sequences of extended length, with a naturalistic hierarchical structure, a rich set of features, and a large number of sequences.

One goal was to maintain a high degree of realism. We aimed to create sequences that would appear natural if presented to human participants, to permit the study of the cognitive processing of events in a way that closely mirrors how this processing occurs in real everyday experiences. Naturalistic stimuli provide a challenging testbed for computational systems with high ecological validity for translating to real-world applications.

A second goal was to create continuous, temporally extended event sequences. Humans segment continuous perceptual input during observation and experience. Yet many existing datasets consist of brief events in isolation. Merging isolated events to form event sequences produces unnatural discontinuities between events, limiting the application of findings to perception in naturalistic environments. We created temporally extended continuous event sequences to study event structure without overt discontinuities between events.

A third goal was to create stimuli with a naturalistic hierarchical structure. Naturally occurring behavior has structure at a range of timescales, producing a hierarchy of events and sub-events (Dickman, 1963). To capture this aspect of activity, we generated scripts for four classes of common everyday activities that naturally consist of a series of steps performed sequentially. One of our chapter types, "making breakfast," is of the class of food preparation sequences that have been employed frequently in both event segmentation (Eisenberg & Zacks, 2016; Swallow et al., 2018) and machine learning studies (Kuehne et al., 2014; Stein & McKenna, 2013; Zhou et al., 2018). We used an algorithm to randomly select actions for the actor to perform (such as "jumping rope"), that comprised sub-chapters (such as cardio exercises), that combined to form complete activity sequences or chapters (such as completing a workout). Thus, we generated activity sequences with three hierarchical levels. Each chapter type was

[1] See Grall and Finn (2021) for a critique of using commercial media as "naturalistic" stimuli.

Table 1 Statistics of relevant event stimulus sets and the META stimulus set (in bold)

	MPII Cooking 2 (Rohrbach et al., 2016)	YouCook2 (Zhou et al., 2018)	50 Salads (Stein & McKenna, 2013)	Breakfast Actions (Kuehne et al., 2014)	Complex and Long Activities Dataset (Tayyub et al., 2017)	Charades- Ego (Sigurdsson et al., 2018)	Charades (Sigurdss on et al., 2016)	robot observing kitchen activities (Duckworth et al.,2016)	CAD- 120 (Koppula et al., 2013)	META Stimulus Set
Total performance duration (hours)	27	175.6	4	17.42	2.24	34.7	82.07	1.39	0.6	**25.77**
Number of sequences	273	2000	50	433	62	4000	9848	493	124	**150**
Average sequence length (minutes)	5.93	5.27	4.8	2.41	2.16	0.52	0.5	0.29	0.29	**10.3**
Number of action classes	59	89	17	10	4485	157	157	11	10	**46**
Number of action instances		1958	966	1989	8846	68536	66500	398	124	**1013**
Number of actors	30	NA	25	52	5	112	267	300	4	**5**
Width x height pixel resolution	1624x1224	various	640x680	320x240	1920x1080	various	various	640x480	320x240	**1920x1080**
Multi-angle	√	√	√	√	√	√	√	√	√	√
Depth	√	√	√	√	√	√	√	√	√	√
Features	Action annotation, hand and pose for subset of frames	Action annotations, objects	Accelerometer data, high and low level activity annotation	Coarse and fine action annotation	Skeletons, action annotation	Action annotation	Action annotation	Skeleton	Skeleton, objects, affordances, sub- and high-level activities	**Skeleton, objects, highlevel activities, event boundaries**

performed multiple times by different actors in different settings, allowing for the study of event schemata using rich naturalistic stimuli (Baldassano et al., 2018; Schank & Abelson, 1975).

A fourth goal was to annotate each captured performance with a rich set of features. When choosing the features to extract and annotate from the recordings, we sought to produce a set of features that would approximate those available to mid-level human visual processing, including biological motion and object interactions (Grill-Spector & Malach, 2004). Prior work has proposed that this level of abstraction may be particularly useful for creating event models, as it provides a smoother representation of meaningful event dynamics above the noise and idiosyncrasies of lower-level vision (Richmond & Zacks, 2017). To this end, we used a depth camera and skeleton tracking algorithm to measure skeletal joint positions in three-dimensional space. Further, we tracked the identities and positions of objects with which the actors interacted and annotated the timings of the high-level event structure. We captured the performances from multiple camera angles to enable the study of how viewpoint affects the processing of naturalistic event sequences (Swallow et al., 2018).

In addition to the features that we created as an approximation of mid-level human vision, we also collected normative event boundaries for all chapters in the META stimulus set. Humans can segment ongoing experiences into events at multiple timescales, and tend to show significant group-level agreement in where they identify boundaries (e.g., Kurby & Zacks, 2011; Zacks et al., 2006). Using an online sample, we collected normative event boundaries for all chapters in the META stimulus set at two grains of segmentation: *coarse*, or the largest meaningful units of activity, and *fine*, or the smallest meaningful units of activity. These data can be used to generate distributions of the likelihood that participants identified boundaries at a coarse and fine grain, providing a psychological grounding of how people tend to segment the continuous event sequences. Thus, the normative segmentation data is useful for developing and evaluating models of how when mid-level visual features can evoke subjective event boundaries.

A final goal of this effort was to produce a corpus large enough to train computational models of event comprehension to near-human-level performance. Producing precisely structured but naturalistic event sequences, we were unable to match the total quantity of stimuli produced through crowdsourcing or collecting videos from online repositories; however, the current corpus is substantially larger than the stimulus sets typically used in psychological research. In total, we created over 25 hours of event sequence performances.

Corpus creation

Overview The META corpus consists of 150 recordings (ranging from 335.45 s to 1309.18 s in duration) of an actor performing one of four types of everyday activity: making breakfast in a kitchen, performing a workout, cleaning a room, or grooming in a bathroom. We refer to each recording as a "chapter." Each chapter was recorded with two high-resolution video cameras and a Kinect 2 sensor (Microsoft, Redmond, WA), which incorporates a video camera and an infrared time-of-flight depth sensor. Five young adult actors (three male, one female, one nonbinary) [2] each performed 10 chapters of three (out of the four total) chapter types, for a total of 30 chapters per actor. See Table S1 for actor demographic information. There was partial data loss for two of the 150 chapters, which is described in more detail in the "Data streams" section. A unique, naturalistically furnished room (kitchen, living room, bedroom, bathroom) was used for each chapter type that an actor performed, for a total of 15 unique locations across the stimulus set.

Sequence generation Naturalistic action has predictable sequential and hierarchical constraints, but also variation. For example, when making toast, the bread must be placed in the toaster before the toaster is started, but fetching a plate to hold the toast could happen before fetching the bread or while the toast is toasting. Moreover, actions can occur in multiple contexts. For example, pouring water into a kettle might happen in the context of making coffee or making tea. To generate action sequences for the actors to perform that realized these characteristics, we created a custom sampling program using a stochastic grammar for naturalistic combinations of actions. In other words, the program randomly selected actions to have unique sequences, with the inclusion of rule-based limitations to avoid unnatural action sequences. Each chapter had three subchapters from which specific actions were sampled (see Fig. 1). For example, the breakfast chapter contained subchapters of preparing main dishes, preparing side items, and preparing beverages. In addition to these chapter-specific subchapters, there was a category of multichapter actions that could appear in any type of chapter. Each chapter was created from sampling two actions for each of the three subchapters (with the exception of sampling one action for the Hair Grooming subchapter because this subchapter had a small number of natural actions that could appear together). In addition, each chapter included one action from the multichapter set of actions, for a total of 6–7 actions per chapter. Subchapter actions

[2] The actors are numbered 1, 2, 3, 4, and 6; actor number 5 left the project early in recording and was replaced with actor number 6.

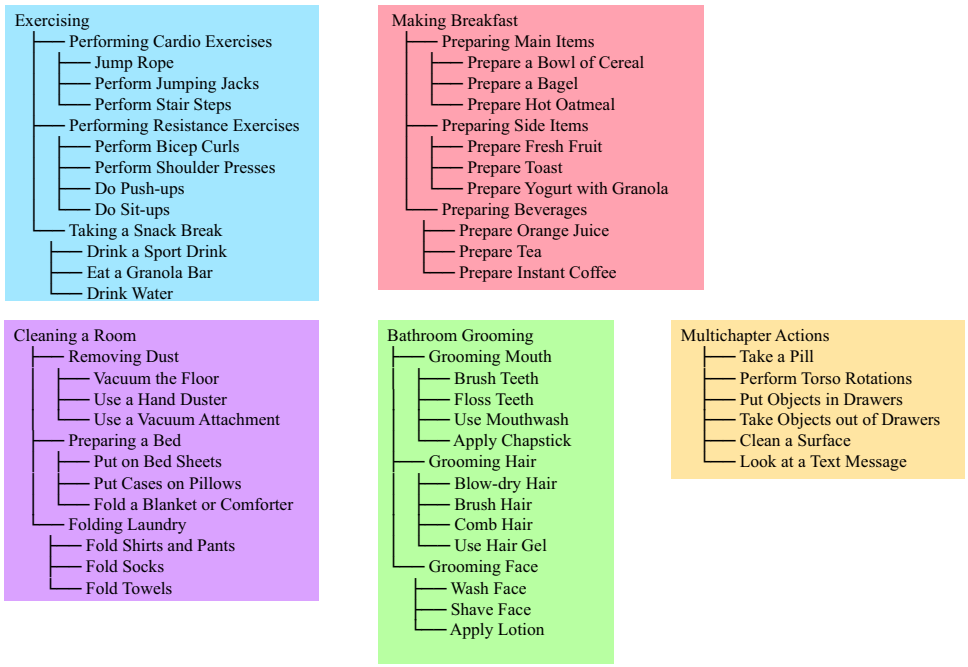


Fig. 1 Hierarchical structure of all actions sampled to generate action sequences. For each chapter type, two actions were sampled for each subchapter (one action for the Grooming Hair subchapter), and one action was sampled from the Multichapter Actions set

were contiguous in time, but the order of actions within a subchapter and the order of subchapters was randomized.

The specific actions that comprised each chapter were selected randomly, with grammatical constraints that we imposed to avoid unrealistic sequences of actions. Specifically, there were *exclusive actions*, two or more actions for which including more than one within the same sequence would appear unnatural, and there were *order constraints,* actions that would appear unnatural to follow other actions. Exclusive actions were making coffee or tea, making cereal or hot oatmeal, and drinking water and drinking a sport drink. Order constraint rules were that applying lotion could not occur before washing face or shaving face, flossing teeth could not appear before brushing teeth, using mouthwash could not appear before brushing teeth or flossing teeth, and applying lip balm could not appear before washing face, shaving face, brushing teeth, flossing teeth, or using mouthwash. To create sequences that fit these constraints, we repeatedly generated random sequences and discarded any that violated the grammatical rules.

Performance Actors memorized the lists of actions to perform for each chapter. In each chapter, the actor entered the room from off screen, sequentially performed the scripted actions without pausing recording or reviewing the scripted actions, then exited off screen. Actors performed each chapter alone, with no interactions with other actors. Each performance was lit with the lighting in the room and two professional halogen lights placed on stands. We used professional lighting both to maintain greater uniformity in lighting across chapters and to improve the tracking of the Kinect sensor. Skeleton tracking was monitored live during the performance, and chapters with poor skeletal tracking were repeated to capture a better take.

Data streams Each performance was captured with three static cameras affixed to tripods. We used two Sony HD Handycam color cameras (model number HDR-PJ260) to capture each performance, each with video parameters of pixel resolution of 1920×1080, MPEG-4 codec, frame rate of 29.97 frames per second, and audio parameters of AC3 codec, sample rate 48,000 Hz. These cameras were placed at about the height of the actor's head and situated at an approximately 90-degree angle from each other, to capture different views of the actors performing the action sequences. In addition to these cameras, a Microsoft Kinect v2 camera was placed near one of the color cameras (see Fig. 2a–c). The Kinect v2 includes a color video camera and an infrared time-of-flight sensor to capture 3D depth data, with the following specifications: color camera resolution = 1920 × 1080, color camera field of view = 84.1° × 53.8°, depth camera resolution = 512 × 424, depth camera field of view = 70.6° × 60°, frame rate = 30 frames per second. Due to technical errors, the Kinect v2 recording was lost for one of the chapters, and one of the color video cameras ended recording mid-performance during another chapter.

Fig. 2 Examples of performance capture and annotations. Each chapter was captured from (**a**, **b**) two HD color cameras and (**c**) a Kinect v2 camera. **d** Skeleton joint positions were estimated in three-dimensional space and mapped to the two-dimensional Kinect color camera. **e** Objects were hand-coded on sampled frames from the videos then tracked on intervening frames with an automated tracking model. Bounding boxes are labeled with object identities and the tracking model's confidence

We used the Kinect v2's built-in algorithms for detecting bodies, detecting the positions of 25 skeletal joints in three-dimensional space, and mapping the three-dimensional joint positions to the two-dimensional space of the Kinect color camera (see Fig. 2d). Live Kinect recording was performed and monitored with the Microsoft Kinect Studio application. Video, audio, depth, and 3D and 2D skeleton data were extracted from playback of the recorded files using custom scripts in the C# language.

Preprocessing of camera data The three cameras were synchronized by matching the sound and visual appearance of the snap of a clapperboard that was visible from all three camera angles. Once the videos were synchronized, the beginning and end of each chapter was trimmed to remove times when crew members were audible or on screen and the pauses before and after the actor began to enter from off screen. Once trimmed, the chapter durations ranged from 335.45 s to 1309.18 s, with $M = 618.36$ s and $SD = 144.80$ s (summarized by chapter type: making breakfast $M = 751.83$ s, $SD = 161.04$ s; exercising $M = 539.66$ s, $SD = 73.20$ s; cleaning a room $M = 607.17$ s, $SD = 98.11$ s; bathroom grooming $M = 537.77$ s, $SD = 115.40$ s), as shown in Fig. 3.

Preprocessing of skeletal data We applied preprocessing steps to the raw skeletal data, with the goal of limiting errors in tracking and aligning skeletons in a common orientation to improve event learning. Although overall the skeleton tracking algorithm accurately captured actors' skeletal poses, the algorithm occasionally committed errors in tracking that we addressed through processing the raw skeletal pose data. At times, phantom skeletons were detected in reflective surfaces and the tracked skeleton could also momentarily leave the actor's body and get stuck on a surface. We coded the timings of these errors and filtered out phantom skeletons or skeletons that left the body for extended periods of time. When the algorithm suffered a momentary lapse in tracking, the body ID of the actor could be replaced by a new body ID. We coded these instances and merged together body IDs that corresponded to the same actor. We also applied 3D translation and rotation to the skeleton joint coordinates to align all skeletons to a common body-centric orientation. Coordinates at each time point were translated to align the origin with the mid-spine joint, then 3D rotation was applied about the y-axis to align the left and right shoulder joints on a common z-plane. The result of this preprocessing was a single skeleton ID for each chapter with off-body times removed and that was synchronized to all cameras, with body-centric 3D coordinates facing forward.

Semi-automated object labeling In each chapter, actors interacted with objects to complete the actions, and understanding these object interactions may aid in forming event representations. To track the positions and identities of the onscreen objects, we used a combination of hand labeling and machine learning object tracking. First, for all three

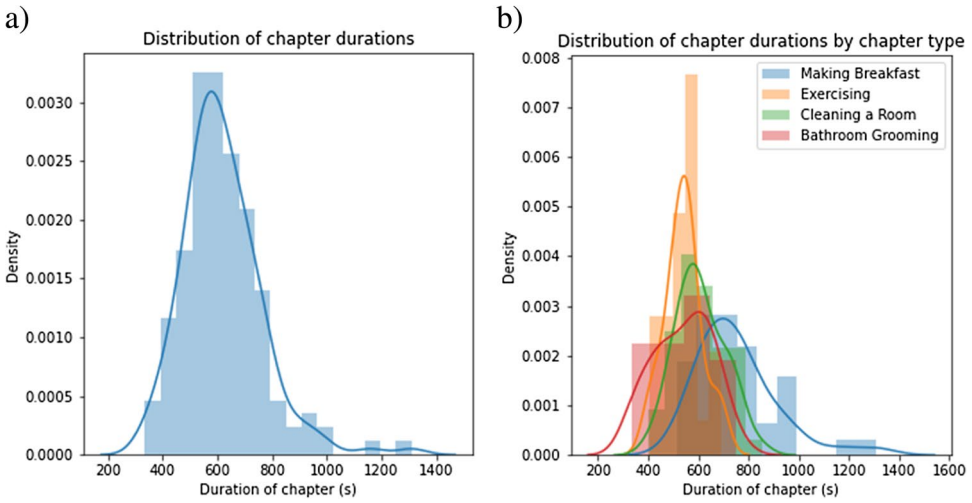


Fig. 3 **a** The distribution of chapter durations. **b** The distribution of chapter durations by chapter type

color camera angles, frames were extracted from each video using FFmpeg (Tomar, 2006), starting at 10 s, and at a rate of every 20 s[3]. Then, using the Python program LabelImg (Tzutalin, 2015), a team of seven coders (with one labeling per image) drew bounding boxes and labels for all objects in the frame. For objects that did not move across successive frames, the locations of bounding boxes were copied over from the previous frame. In total, there were 16,737 ground truth frames labeled, with 171,006 total instances across all frames of 113 classes of objects (see Table S2 for a list of the labeled object classes).

Then, using the subset of hand-labeled frames, the positions of objects were interpolated across the intervening unlabeled frames using a tracker called the Siamese region proposal network (SiamRPN, B. Li et al., 2018), implemented using the Python Surrogate Optimization Toolbox (pySOT) of the Python library. Specifically, from hand-labeled frames, one SiamRPN model is instantiated for each object. For each 10-second interval, we tracked positions of objects both forward and backward, and each tracker was terminated when its confidence fell below a threshold. We then ran the Hungarian algorithm (Kuhn, 1955) to match forward tracks and backward tracks. We repeated forward tracking from the results of matching, and backward tracking from the following extracted hand-label frame. This process produced labels and locations for all interactive objects for all frames of the videos in the META stimulus set (see Fig. 2e).

[3] For the Breakfast and Exercise chapters of Actor 1 (20 chapters), the frames were extracted starting at 5 s and every 10 s). After extracting and annotating these frames, we moved to the sparser frame labeling rate because it produced acceptable tracking results in a shorter amount of time required for labeling.

Semantic embeddings of object interactions Language models trained on large natural datasets can provide vector embeddings of the semantic meaning of object labels. Here we provide one such embedded representation, though many other options exist. We employed the GloVe model (Pennington et al., 2014), trained on the Wikipedia 2014 and the Gigaword Fifth Edition (Gigaword 5) corpora, comprising Wikipedia articles and newspaper articles, to produce 50-dimensional vector embeddings. We used these embeddings in two ways: to generate *scene* representations and to generate *nearby-object* representations. Scene vectors were computed at each time point as 50-dimensional equal-weighted averages of the vector representations of all objects present in the scene. Nearby-object vectors were computed as weighted averages, scaled by the inverse of the three-dimensional Euclidean distance between the actor's right hand and the depth of the object in the z dimension as measured with the Kinect depth camera. Thus, the scene vector represents a 50-dimensional general semantic embedding of objects currently present in the environment, and the nearby-object vector represents a 50-dimensional semantic embedding of objects currently near the actor's hand.

Dimension reduction of scene vectors We performed principal component analysis to reduce the dimensionality of the combination of features, both to reduce the computational complexity of the scene vectors and to reduce collinearity between features. We examined scree plots of principal components computed separately for skeleton and semantic features, and chose a number of dimensions that would preserve most of the original variance from the full set of features. We reduced the total features to 30 dimensions (14 skeleton dimensions, 13 semantic dimensions, 1 for object appearances, 1 for object disappearances, and 1 for the correlation

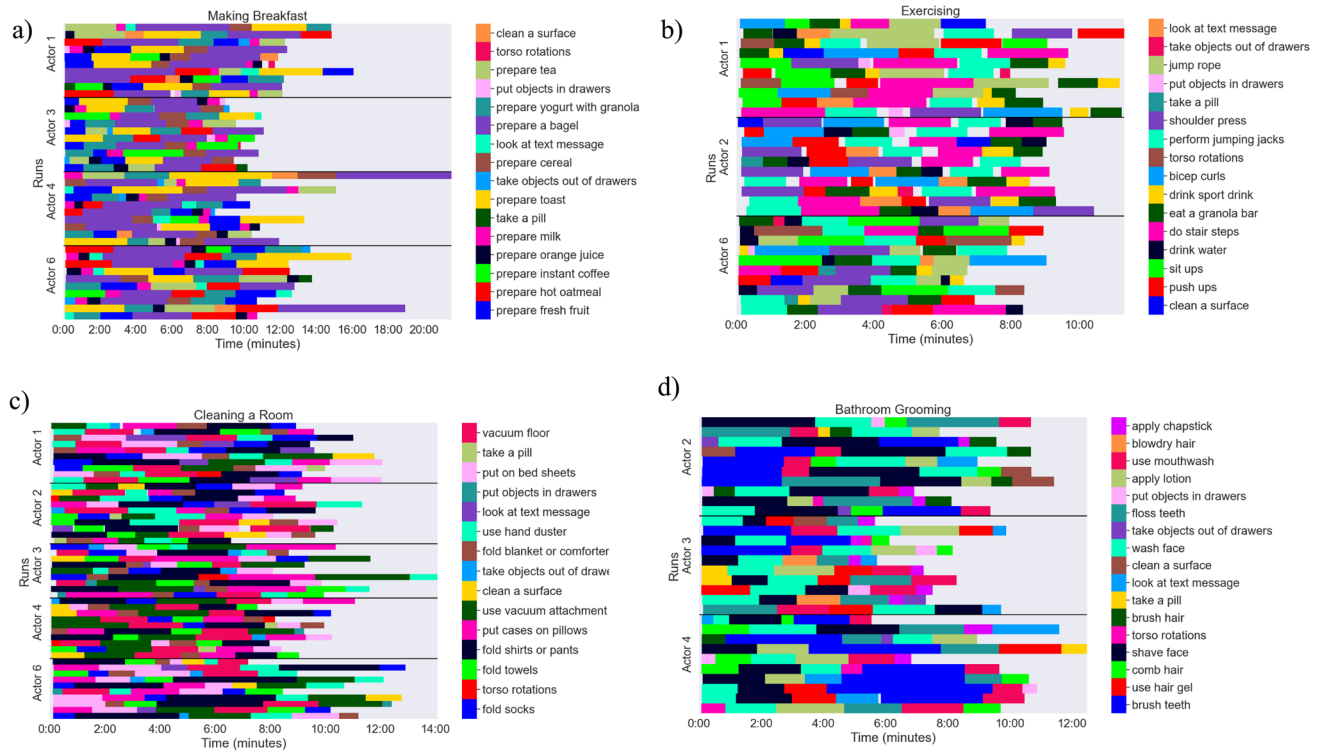


Fig. 4 Annotations of start and stop times for scripted actions by chapter type: **a** making breakfast, **b** exercising, **c** cleaning a room, **d** bathroom grooming

in pixel luminance between each pair of successive frames. In total, this procedure preserved 76 percent of the original variance (see supplementary materials for more details; PCA files are included on the Open Science Framework [OSF] repository: https://osf.io/q7yu2/). We present the 30-dimensional scene vectors as one computationally tractable representation of the activity sequences that approximates the level of abstraction afforded by mid-level human vision, and represents information that may be useful for a computational model to learn event dynamics[4].

Annotation of scripted actions A human rater coded the start and end times of each of the scripted actions that the actors were asked to perform. The annotated actions had a median duration of 76.61 s with a standard deviation of 54.63 s. Figure 4 shows the annotated action timings for each of the four chapter types.

The present study

We conducted a preregistered experiment to collect normative data representing when people perceive event boundaries in the recorded activities (https://doi.org/10.17605/OSF.IO/A9GUZ). For each chapter, participants marked event boundaries at coarse or fine grains of segmentation. We predicted that for all videos, the fine segmentation condition would produce a larger number of boundaries than the coarse segmentation condition (i.e., a shorter mean event duration), and that there would be higher than chance agreement across participants in the locations of event boundaries for both fine and coarse segmentation.

Method

Participants We recruited a total of 3090 participants: 2956 through Amazon Mechanical Turk and 134 through the Washington University in St. Louis Psychological & Brain Sciences participant pool. Participants self-reported their age (median = 32 years, SD = 36.87 years), gender (1136 female, 1196 male, and 13 other), race (57 American Indian or Alaska Native, 263 Asian, 625 Black or African American, 2190 White), and ethnicity (647 Hispanic or Latino). Online experiments were managed using the CloudResearch

[4] Note that in reducing the set of features to 30 dimensions, our purpose was to orthogonalize the features and reduce the computational cost. We do not make any claim that mid-level human vision is best captured in a 30-dimensional space.

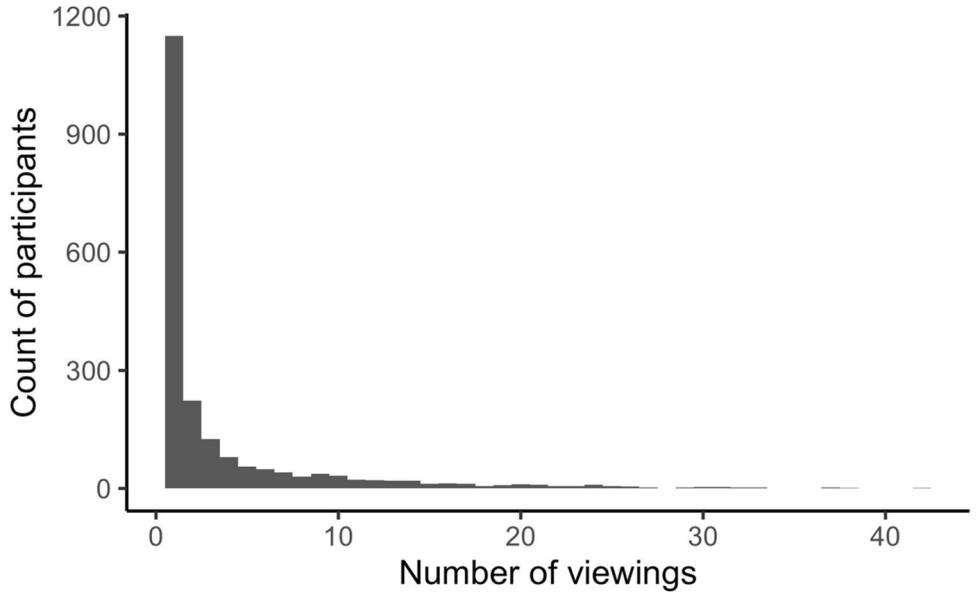
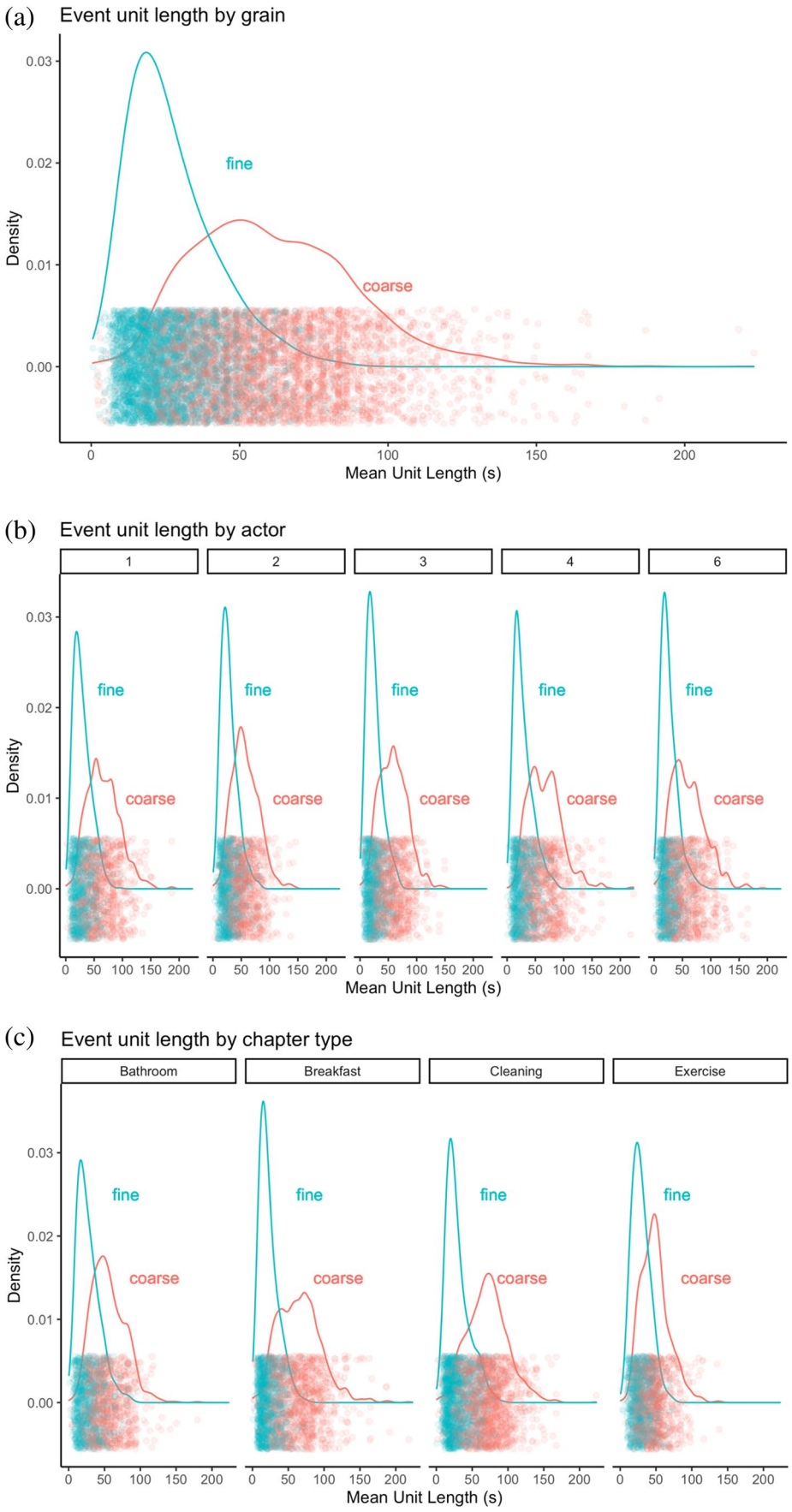


Fig. 5 Distribution of films segmented per participant

platform (cloudresearch.com). Participants were compensated with $3 or course credit for each session of segmenting a single video. The following inclusion criteria were used when recruiting through Amazon Mechanical Turk to maximize data quality: workers needed to have an approval rate of at least 95 percent, to have completed at least 500 jobs, and to be accessing the internet from within the United States. A catch question was also included on the demographic questionnaire that participants completed at the end of the session, which asked the participant to select an answer for a multiple-choice question. If participants answered the catch question incorrectly, their data were excluded from analysis. Participants who met the inclusion criteria were permitted to segment multiple videos, and the number of videos segmented per participant ranged from 1 to 44, with a median value of 1 (see Fig. 5 for the distribution of the number of videos each participant segmented). Each time a participant returned to the experiment, a video at their assigned grain of segmentation was randomly selected. We continued to collect data until we obtained 30 segmentations from unique participants for each grain for each video.

Procedure New participants were randomly assigned a segmentation grain of "coarse" or "fine." Following the procedure used by Newtson (1973), participants were given these instructions (with words that differed between the coarse and fine conditions in brackets): "We would like you to watch a movie of an actor performing everyday activities. As you watch, we would like you to mark off each time you judge that the [LARGEST]/[SMALLEST] meaningful unit of activity has ended and another has begun. To mark off a boundary between two units of activity, press the SPACEBAR. Please be careful to press the button as close to the end of the unit as possible. Do not press the button in the middle of a unit. Before starting the main task, you will mark the boundaries in a brief practice video so we can make sure that your performance matches the typical viewer. Press the 'Play' button when you are ready to begin." Participants first segmented a practice video with a duration of 2 min 35 s in which a man constructs a toy boat using interlocking building blocks. If participants marked fewer than three boundaries in the coarse condition or fewer than six boundaries in the fine condition, they received the message: "People typically identify [3–4]/[6–8] units during this practice movie. You identified [X] units during this movie. We will continue to the main task, but please attempt to identify units more frequently in the next movie. Remember, press the SPACEBAR whenever you believe that one [LARGE]/[SMALL] meaningful unit of activity has ended and another begins. Click the button below to continue." Then participants segmented one of the videos at their assigned grain. Videos were preloaded to the participant's computer before playing to prevent buffering lags during playback, and controls were disabled using JavaScript to prevent participants from pausing or seeking to other times while segmenting the videos. After segmenting the video, participants completed a demographic questionnaire that included a catch question as an attention check. Participants were permitted to repeat participation if they met the inclusion criteria. Repeat participants did not recomplete the demographic questionnaire again. If participants returned within a week of their prior


(a)
Event unit length by grain
Density
0.03
0.02
0.01
0.00
fine
coarse
0
50
100
150
200
Mean Unit Length (s)
(b)
Event unit length by actor
1
2
3
4
6
Density
fine
coarse
Mean Unit Length (s)
(c)
Event unit length by chapter type
Bathroom
Breakfast
Cleaning
Exercise
Density
fine
coarse
Mean Unit Length (s)

◂**Fig. 6** **a** Mean event unit lengths were longer for the coarse segmentation condition than the fine segmentation condition. Each point depicts the mean unit length for a single session of a participant. The difference between coarse and fine boundary unit lengths was also observed across **b** actors and **c** chapter types

session, they did not have to recomplete the practice segmentation portion of the experiment.

Analysis Data were processed using custom R scripts. From a set of 10,560 sessions in the raw dataset, 781 sessions were filtered from participants who answered the catch question incorrectly. Then, 306 sessions were filtered due to a lack of recorded button presses during the experiment. In addition to the exclusion criteria set in our preregistration, we applied two more steps to reduce the noise of data collected through an online sample. We measured timestamps recorded at the start and end of playback and compared this duration to the duration of the videos. If the playback timestamps deviated from the movie duration by more than 2.5 s, we excluded that viewing from analysis. This criterion removed another 228 sessions. To further constrain the segmentation data, we excluded participants who identified fewer than one third or more than three times the median number of boundaries for the segmentation grain of each film, removing an additional 1122 sessions. It was discovered that some participants were mistakenly assigned the same video multiple times. Removing duplicate video sessions filtered an additional 170 sessions. Finally, there were cases in which despite assigning participants to a single grain, they completed sessions at the unassigned grain. Removing the sessions at the unassigned grain filtered an additional 21 sessions. After filtering based on these criteria, the resulting dataset contained 7931 sessions. Applying these additional exclusion criteria did not change the pattern or significance of results (see supplementary materials for results using original exclusion criteria: https://osf.io/p56gh/).

In order to confirm our manipulation of the fine and coarse grain instructions, we tested for differences in unit length between coarse and fine segmentation using a linear mixed model implemented with the lmer function of the lme4 package in R (Bates et al., 2015)[5]. We tested the fixed effect of segmentation grain (Condition) on mean unit length (Mean.Length) with random effects of video (Movie) and participant (workerId). In Wilkinson-Rogers notation (Wilkinson & Rogers, 1973), the formula for this model is: Mean. Length ~ Condition + (1|Movie) + (1|workerId). We compared this model to a model without Condition to see whether including segmentation condition provided a better model fit for predicting mean unit length. We also compared the Condition model with models that included factors for Actor and Chapter Type to see whether these factors improved model fit.

We measured agreement in segmentation between participants separately for each combination of chapter and segmentation grain. We binned the segmentation data into 1-s bins, and then computed the correlation coefficients between a single participant's binned segmentation and the proportional binned segmentation of all other participants who segmented that chapter at that grain. This process was repeated with each participant as the left-out binned segmentation for comparison. Because participants identified different numbers of boundaries, the range of possible correlation values will be different across participants. This makes it difficult, for example, to compare a participant who identified a small number of boundaries that aligned with popular segmentation bins to a participant that identified a large number of boundaries with more inconsistent alignment with popular segmentation bins. We employed a method developed by Kurby and Zacks (2011) to correct for the differences in possible correlation values at the given number of boundaries. For each number of boundaries, we computed the highest possible correlation value (i.e., if the boundaries occurred at the highest-agreement time bins) and the lowest possible correlation value (i.e., if the boundaries occurred at the lowest-agreement time bins). We then scaled the correlation coefficients from zero to one based on the minimum and maximum correlations possible for the given number of boundaries and the group distribution. To test whether the observed level of agreement was better than would be predicted by chance, we calculated the mean scaled correlation for coarse and fine segmentation, and created null distributions by sampling with replacement 100 times the number of participants per cell and shuffling the unit lengths of marked events. We computed z statistics for the observed agreement relative to the null distribution and converted the z scores to p-values using a normal distribution to test for significance.

Results and discussion

Participants identified events with a shorter mean duration at a fine grain than at a coarse grain As expected, the mean event unit length for the fine condition (median = 23.45 s, SD = 14.81 s) was shorter than for the coarse condition (median = 58.60 s, SD = 27.60 s), replicating prior research (Kurby & Zacks, 2011; Swallow et al., 2018; Zacks, Tversky,

[5] Testing the skewness of the event length distribution revealed a small positive skew (skewness coefficient of 1.03). As this level of skew does not exceed cutoffs typically used for approximating a normal distribution, e.g., skew < 1.5 (Tabachnick et al., 2019), we do not transform the data in the reported results. Repeating the analysis with log-transformed event lengths did not change the pattern or significance level of the results.

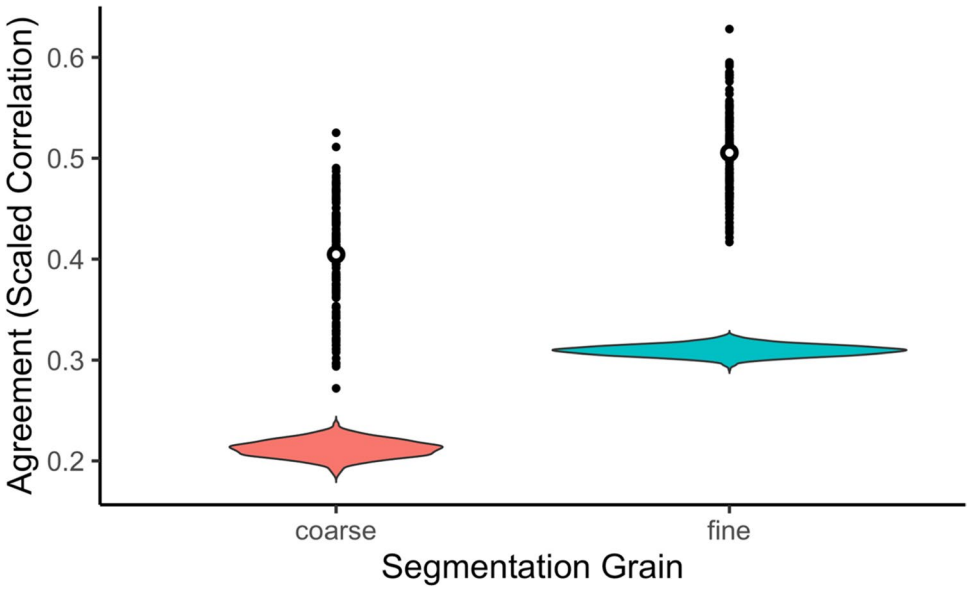


Fig. 7 For both fine- and coarse-grained segmentation, agreement between participants was significantly better than chance. Violin plots depict null distributions generated from repeatedly shuffling the order of event unit lengths for each video and computing agreement. Points depict the observed agreement (as scaled correlation coefficients) between participants for each video, with the mean observed agreement across all videos shown as rings with white centers

et al., 2001; Zacks & Swallow, 2007). Including segmentation condition when modeling the mean unit length significantly fit the data better than omitting segmentation condition (*with segmentation condition AIC* = 68772, *without segmentation condition AIC* = 69839, $\chi^2 = 1069.4$, $df = 1$, $p <$ .001). Adding the actor to the model with segmentation condition did not significantly improve the model fit ($p = .45$), but adding chapter type did (*with segmentation condition and chapter type AIC* = 68712, *with segmentation condition AIC* = 68772, $\chi^2 = 65.574$, $df = 3$, $p < .001$). See Fig. 6 for the comparison of unit lengths by segmentation condition, chapter type, and actor. In comparison to the distribution of annotated scripted action durations, the distribution of coarse event unit lengths overlaps, with a shorter median unit length, indicating that the largest subjective units of activity that participants identified tended to be smaller than the scripted actions performed by the actors.

Participants showed higher than chance level agreement for fine and coarse segmentation Participants showed moderate agreement in where they identified boundaries, and agreement was better for fine segmentation (mean scaled $r = 0.51$) than for coarse segmentation (mean scaled $r = 0.41$). This level of inter-subject agreement was significantly better than the chance level (*fine* and *coarse* $p < .001$) predicted by the null distribution generated from shuffling the event unit lengths (see Fig. 7 for agreement values and Fig. 8 for examples of coarse and fine event boundaries).

The results of the segmentation experiment provided normative boundaries at both a fine and a coarse grain of segmentation. For both grains, agreement between participants was significantly greater than agreement predicted by chance. This level of agreement was similar to the level of agreement reported in other experiments with videos of everyday events (Kurby & Zacks, 2011; Zacks et al., 2006), and was significantly higher than chance, even when using a conservative method for permuting the null distribution that preserves the unit lengths of the observed human event boundaries. When using these segmentation data for applications in the study of events, group-binned boundaries may be thresholded to produce discrete boundaries of higher agreement among participants, or the continuous distribution of participant boundaries over time may be used as a measure of boundary likelihood (Ben-Yakov & Henson, 2018).

Conclusion

Here we report a large, naturalistic corpus of recordings of everyday activities, using video camera and depth image recording with dense annotation. We have included a set of features and annotations of the META stimulus set that we believe are useful in the comprehension of unfolding events. In sharing these materials publicly, it is our hope that others will create additional features and annotations to improve upon those that we have provided, expanding the usefulness of this tool. For example, to complement the high-level scripted action annotations, it would be useful to code timings and descriptions of fine-grained actions performed by

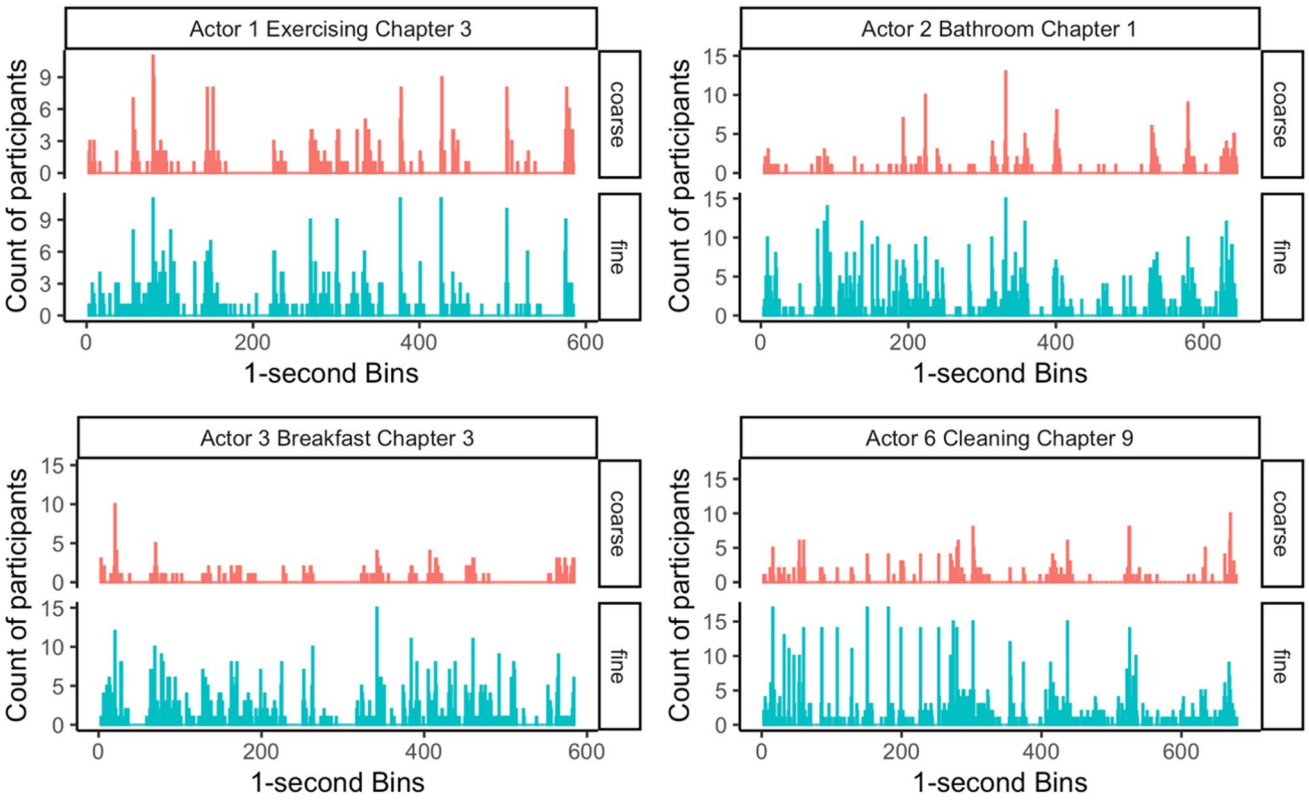


Fig. 8 Group coarse and fine event boundaries for four example videos in the META stimulus set. Individual participant boundaries were binned into 1-s bins and plotted as the number of participants who identified a boundary in each bin

the actors, as well as precise contacts and transformations performed on objects. By creating event sequences that are more realistic, of a longer duration per chapter, and richly annotated, we aim to advance the capabilities of tools available for investigating the continuous dynamic events. We hope that the META stimulus set will be useful for a broad range of applications for researchers who study dynamic complex events.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01980-8.

Acknowledgements This work was supported by a grant from the Office of Naval Research: [Grant Number N00014-17-1]. We thank Garrett Cunningham, Sarah Hale, Ryan Kahle, Duy Pham, and Chong Wang for performing the chapters used in the stimulus set. We thank Cory Fox, Emma Lavetter-Keidan, Sierra Revels, Matt Steinhaus, and Grace Zhou for assistance in creating annotations. We thank Maverick Smith for helpful comments on an earlier draft of the manuscript.

Data availability The materials of the META stimulus set, including videos, depth data, skeleton pose data, tracked object positions, annotated action timings, anonymized data from the normative event boundaries experiment, and processing scripts are available through a repository on the Open Science Framework at https://osf.io/3embr/. The normative event boundaries experiment was preregistered, and the preregistration can be found here: https://osf.io/a9guz.

References

Aakur, S. N., & Sarkar, S. (2019). A perceptual prediction framework for self supervised event segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1197–1206.

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*, 715–729.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning & Verbal Behavior, 17*(1), 1–12.

Bailey, H. R., Kurby, C. A., Giovannetti, T., & Zacks, J. M. (2013). Action perception predicts action performance. *Neuropsychologia, 51*(11), 2294–2304. https://doi.org/10.1016/j.neuropsychologia.2013.06.022

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 http://www.sciencedirect.com/science/article/pii/S0896627317305937

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience, 38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*(3), 708–717.

Barbey, A., Krueger, F., & Grafman, J. (2009). Structured event complexes in the medial prefrontal cortex support counterfactual representations for future planning. *Philosophical Transactions of the Royal Society B-Biological Science, 364*(1521), 1291–1300. https://doi.org/10.1098/rstb.2008.0315

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

Ben-Yakov, A., & Henson, R. (2018). The hippocampal film-editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience, 0524–18*. https://doi.org/10.1523/JNEUROSCI.0524-18.2018

Bonasia, K., Sekeres, M. J., Gilboa, A., Grady, C. L., Winocur, G., & Moscovitch, M. (2018). Prior knowledge modulates the neural substrates of encoding and retrieving naturalistic events at short and long delays. *Neurobiology of Learning and Memory, 153*, 26–39. https://doi.org/10.1016/j.nlm.2018.02.017

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177–220.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 717–726. https://doi.org/10.1016/S0022-5371(72)80006-9

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A., & Otte, S. (2019). Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks, 117*, 135–144.

Dickman, H. R. (1963). The perception of behavioral units. In R. G. Barker (Ed.), *The stream of behavior* (pp. 23–41). Appleton-Century-Crofts.

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

Duckworth, P., Alomari, M., Gatsoulis, Y., Hogg, D. C., & Cohn, A. G. (2016). Unsupervised activity recognition using latent semantic analysis on a mobile robot. *IOS Press Proceedings, 285*, 1062–1070.

Eisenberg, M. L., Sargent, J. Q., & Zacks, J. M. (2016). Posttraumatic stress and the comprehension of everyday activity. *Collabra, 2*(1).

Eisenberg, M. L., & Zacks, J. M. (2016). Ambient and focal visual processing of naturalistic activity. *Journal of Vision, 16*(2), 5. https://doi.org/10.1167/16.2.5

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

Fivush, R. (1997). Event memory in early childhood. In N. Cowan (Ed.), *the development of memory in childhood* (pp. 139–161). Psychology press/Erlbaum (UK) Taylor & Francis; psyh.

Franklin, N., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2019). Structured event memory: A neuro-symbolic model of event cognition. *BioRxiv, 541607*. https://doi.org/10.1101/541607

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Graesser, A. C., & Nakamura, G. V. (1982). *The impact of a schema on comprehension and memory: Vol. the psychology of learning and motivation, Vol. 16* (G. H. Bower, Ed.; pp. 59–109). Academic press.

Grafman, J., Sirigu, A., Spector, L., & Hendler, J. (1993). Damage to the prefrontal cortex leads to decomposition of structured event complexes. *Journal of Head Trauma and Rehabilitation, 8*(1), 73–87.

Grall, C., & Finn, E. S. (2021). The 'naturalistic' fallacy: Leveraging the power of media to drive cognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/c8z9t

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience, 27*(1), 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience, 28*(10), 2539–2550 http://www.jneurosci.org/cgi/content/abstract/28/10/2539

Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology, 45*(2), 575.

Koppula, H. S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research, 32*(8), 951–970.

Kuehne, H., Arslan, A., & Serre, T. (2014). The language of actions: Recovering the syntax and semantics of goal-directed human activities. *IEEE Conference on Computer Vision and Pattern Recognition, 2014*, 780–787. https://doi.org/10.1109/CVPR.2014.105

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. *International Conference on Computer Vision, 2011*, 2556–2563.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97. https://doi.org/10.1002/nav.3800020109

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition, 39*(1), 75–91. https://doi.org/10.3758/s13421-010-0027-2

Kurby, C. A., & Zacks, J. M. (2018). Preserved neural event segmentation in healthy older adults. *Psychology and Aging, 33*(2), 232–245. https://doi.org/10.1037/pag0000226

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, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with Siamese region proposal network. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018*, 8971–8980. https://doi.org/10.1109/CVPR.2018.00935

Li, Y.-L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.-S., Wang, Y., & Lu, C. (2019). Transferable interactiveness knowledge for human-object interaction detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3585–3594.

Liang, M., & Hu, X. (2015). *Recurrent convolutional neural network for object recognition* (pp. 3367–3375) https://openaccess.thecvf.com/content_cvpr_2015/html/Liang_Recurrent_Convolutional_Neural_2015_CVPR_paper.html

Lichtenstein, E. D., & Brewer, W. F. (1980). Memory for goal-directed events. *Cognitive Psychology, 12*, 412–445.

Magliano, J., Kopp, K., McNerney, M. W., Radvansky, G. A., & Zacks, J. M. (2012). Aging and perceived event structure as a function of modality. *Aging, Neuropsychology, and Cognition, 19*(1–2), 264–282.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, 2929–2936.

Michelmann, S., Hasson, U., & Norman, K. (2021). Event boundaries are steppingstones for memory retrieval. *PsyArXiv.* 10.31234/osf.io/k8j94.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., & Vondrick, C. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 502–508.

Nelson, K. (1986). Event knowledge and cognitive development. In *event knowledge: Structure and function in development: Vol. event knowledge: Structure and function in development* (pp. 1–19). Lawrence Erlbaum associates.

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

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford University Press.

Richmond, L. L., Gold, D. A., & Zacks, J. M. (2017). Event perception: Translations and applications. *Journal of Applied Research in Memory and Cognition, 6*(2), 111–120. https://doi.org/10.1016/j.jarmac.2016.11.002

Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences, 21*(12), 962–980. https://doi.org/10.1016/j.tics.2017.08.005

Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., & Schiele, B. (2016). Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision, 119*(3), 346–373.

Roll, E. E., Giovannetti, T., Libon, D. J., & Eppig, J. (2017). Everyday task knowledge and everyday function in dementia. *Journal of Neuropsychology*. https://doi.org/10.1111/jnp.12135

Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review, 122*(1), 1–23. https://doi.org/10.1037/a0037907

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education: Vol. theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education (pp. 33–58)* L. Erlbaum Associates.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition, 129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8*(1), 113–128.

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences, 17*, 41–50. https://doi.org/10.1016/j.cobeha.2017.06.002

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. *IJCAI*, 151–157.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience, 16*(4), 486–492. https://doi.org/10.1038/nn.3331

Sherrill, A. M., & Magliano, J. P. (2017). Psychopathology applications of event perception basic research: Anticipating the road ahead using posttraumatic stress disorder as an example. *Journal of Applied Research in Memory and Cognition, 6*(2), 144–149. https://doi.org/10.1016/j.jarmac.2017.01.004

Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science, 13*(1), 106–127.

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., & Alahari, K. (2018). Actor and observer: Joint modeling of first and third-person videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7396–7404.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision – ECCV 2016* (pp. 510–526). Springer International Publishing.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, *27*. https://papers.nips.cc/paper/2014/hash/00ec53c4682d36f5c4359f4ae7bd7ba1-Abstract.html

Stein, S., & McKenna, S. J. (2013). Combining embedded accelerometers with computer vision for recognizing food preparation activities. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 729–738. https://doi.org/10.1145/2493432.2493482

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

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics*. (Seventh edition). Pearson.

Tayyub, J., Hawasly, M., Hogg, D. C., & Cohn, A. G. (2017). CLAD: A complex and long activities dataset with rich crowdsourced annotations. *ArXiv Preprint* ArXiv:1709.03456.

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal, 146* https://www.linuxjournal.com/article/8517

Tzutalin. (2015). *LabelImg* (Git code) [Computer software]. https://github.com/tzutalin/labelImg

Wang, X., & Ji, Q. (2015). Video event recognition with deep hierarchical context model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015*, 4418–4427. https://doi.org/10.1109/CVPR.2015.7299071

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics, 22*(3), 392. https://doi.org/10.2307/2346786

Yao, B., & Fei-Fei, L. (2010). Modeling mutual context of object and human pose in human-object interaction activities. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010*, 17–24.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001a). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*(6), 651–655.

Zacks, J. M., & Sargent, J. Q. (2010). Event perception: A theory and its application to clinical neuroscience. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 253–299). Elsevier Academic Press.

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.

Zacks, J. M., Tversky, B., & Iyer, G. (2001b). 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

Zalla, T., Labruyère, N., & Georgieff, N. (2013). Perceiving goals and actions in individuals with autism Spectrum disorders. *Journal of Autism and Developmental Disorders, 43*(10), 2353–2365. https://doi.org/10.1007/s10803-013-1784-0

Zalla, T., Verlut, I., Franck, N., Puzenat, D., & Sirigu, A. (2004). Perception of dynamic action in patients with schizophrenia. *Psychiatry Research, 128*(1), 39–51. https://doi.org/10.1016/j.psychres.2003.12.026

Zhang, Y., Qiu, Z., Yao, T., Liu, D., & Mei, T. (2018). Fully convolutional adaptation networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6810–6818.

Zhou, L., Xu, C., & Corso, J. J. (2018). Towards automatic learning of procedures from web instructional videos. *Thirty-Second AAAI Conference on Artificial Intelligence*.