Event segmentation in continuous, naturalistic videos from model-based, data-driven, and human perspectives

Alberto Mariola[1,2,3], Zafeirios Fountas[5], Lionel Barnett[2,3], Warrick Roseboom[2,3,4]

[1] Sussex Neuroscience, School of Life Sciences, University of Sussex, Brighton, UK

[2] Sussex Centre for Consciousness Science, School of Engineering and Informatics, University of Sussex, Brighton, UK

[3] School of Engineering and Informatics, University of Sussex, Brighton, UK

[4] School of Psychology, University of Sussex, Brighton, UK

[5] Huawei 2012 Laboratories, UK

Correspondence

A.Mariola@sussex.ac.uk

Key words

event segmentation; EEG; time perception; memory

Abstract

Human experience is characterised by an apparently continuous stream of information. However, when recalling our past, we commonly experience somewhat discrete episodes. Event segmentation has been suggested to provide a basis for episodic memory formation and recognition/recall. Previous work has suggested that event boundaries are reflected in shifts in stable patterns of neural activity across processing hierarchies, and that these boundaries can be detected using neuroimaging techniques. From a different direction, a recent series of studies in the domain of human time perception have shown that tracking event boundaries as "surprise" in network activity - in human perceptual processing or neural network proxies - successfully describes subjective reports of time on the scale of seconds to minutes.

This project aims to understand how these methods of obtaining “event boundaries” perform relative to human-provided event annotations, in situations of life-like, naturalistic sensory stimulation. We compared the boundaries estimated from our neural-network-based model of time perception and a data-driven event segmentation approach (Greedy State Boundary Search on EEG data), against annotations from 131 human raters using two different inferential processes. In both methods, event boundaries from our model-based and the EEG data-driven approach performed similarly well, and were both indistinguishable from human provided boundaries in at least ~89% of cases examined. A random boundary generator was similar to human raters in only approximately 13% of cases. This work demonstrates the potential to reconcile the fields of time perception and episodic memory through a common foundation in the processes underlying event segmentation.

Introduction

Our perceptual experience typically unfolds continuously and without any perceived interruption. Conversely, when we try to remember something, what used to feel as a dynamic flow turns into a series of discrete events. While not explicit, this segmentation naturally takes place whenever we experience something (Newtson, 1973). According to Event Segmentation Theory (EST - Zacks et al., 2007), events correspond to moments of time taking place in a specific context that are characterised by a beginning and an end (Zacks & Tversky, 2001). EST assumes that humans build internal event models, namely multimodal representations of *"what is happening right now"* that can be used to anticipate changes in the environment and facilitate future behaviour (Kurby & Zacks, 2008; Radvansky & Zacks, 2014; Richmond & Zacks, 2017). In this context, transitions between events - event boundaries - have been proposed to be generated as a result of violations of internal model predictions of the upcoming sensory information, referred to as prediction errors (Zacks et al., 2007, 2011). While a debate about the sufficiency of prediction errors to account for all aspects of defining an event boundary is ongoing (Baldwin & Kosie, 2021; Bromis et al., 2022; Pettijohn & Radvansky, 2016; Schapiro et al., 2013; Shin & DuBrow, 2021; Zakharov et al., 2022), empirical evidence is emerging to support the idea that a change in the expected properties of unfolding experience is a fundamental mechanism underpinning event segmentation (Antony et al., 2021; Kumar et al., 2022; Reynolds et al., 2007; Zacks et al., 2007, 2011).

At the neural level, fMRI studies have found that the generation of event boundaries is associated with evoked responses in a variety of regions including the extrastriate motion complex (MT+), frontal eye fields (FEF), anterior and superior temporal gyri and precuneus (Speer et al., 2003, 2007; Zacks et al., 2001) and with increased hippocampal activity happening in concomitance with moments of transition between events (Ben-Yakov & Henson, 2018). More recently, these findings have been extended and refined by the application of data-driven models such as Hidden-Markov Models (HMM - Baldassano et al., 2017) and the Greedy State Boundary Search algorithm (GSBS - Geerligs et al., 2021) which are able to identify timepoints in fMRI activity that underlie the transition between stable patterns of neural activity mirroring a nested temporal cortical hierarchy of events. Coherent with previous findings on the temporal dynamics of information integration across different brain areas (e.g., Honey et al., 2012; Kiebel et al., 2008; Lerner et al., 2011; Stephens et al., 2013), shorter periods of pattern stability are predominant in primary sensory regions, while longer timescales of pattern stability can be seen along the posterior-frontal axis, consistent with being related to events that can span many seconds (Geerligs et al., 2022). An extensive literature shows that event segmentation is an adaptive mechanism as it promotes successful memory retrieval, generalisation and decision making (e.g., see Bird, 2020; Shin & DuBrow, 2021 for review). For instance, event segmentation has been suggested to explain unique variance in event memory across lifespan (Sargent et al., 2013).

The empirical literature on event segmentation is vast. Many of the studies investigating event segmentation have employed lists of images or words interleaved with artificial boundaries (e.g., auditory tones: Clewett et al., 2020; change of context/task: DuBrow & Davachi, 2013, 2014; coloured frames: Heusser et al., 2018; Pu et al., 2022). Some researchers have used more naturalistic stimuli, such as clips of very specific everyday activities (e.g., Kurby & Zacks, 2011; Zacks et al., 2006, 2011; Zacks & Tversky, 2001). More recently, in several notable cases, researchers have used excerpts from edited/deliberately constructed material such as audio narratives, movies and/or tv shows (e.g., Baldassano et al., 2017; Chen et al., 2017; Silva et al., 2019) or have taken advantage of the highly structured nature of naturalistic events (e.g. a basketball game; Antony et al., 2021). While clearly being more ecologically valid than the very simple stimuli that are often used, media stimuli can also be problematic because event boundaries are not to be "generated" but rather detected among the strong perceptual transitions and/or semantic shifts in the narrative that are purposely included by directors (e.g. see Grall & Finn, 2022 for a general discussion about the use of media stimuli in neuroscience). The aforementioned practice might then lead to circular inference: investigating the capacity of segmenting a scene, when the events are already explicitly "given" to participants.

In contrast with previous works directed specifically at event segmentation, a recent series of studies has approached the relationship between memory and event boundaries from a different perspective - making models of human time perception (Fountas et al., 2022; Roseboom et al., 2019, 2022; Sherman et al., 2022). In the first study of this series, Roseboom et al. (2019) asked human participants to estimate the duration of silent video clips recorded in naturalistic, everyday scenarios (walking in a park and/or on the street, people working in an office, etc.). To produce a model of these estimates of time, the same video clips human participants had viewed were input frame-by-frame to a pre-trained deep neural network (AlexNet - Krizhevsky et al., 2017). Euclidean distance in network activity for each network layer was calculated between successive inputs and then categorised as a salient event by means of a decaying threshold. In order to produce estimates in seconds to compare against human provided reports, the number of events accumulated across each epoch was regressed against physical video duration. Using this method, it was possible to reproduce several of the biases found in human duration reports regarding the same naturalistic stimuli simply by tracking moments of relatively large change in the dynamics of a visual classification network.

In a subsequent study, Sherman and colleagues (2022) had participants do the same task as in Roseboom et al. (2019) while brain imaging was acquired (fMRI BOLD). Using a model-based analysis of the neuroimaging data based on the model from Roseboom et al. (2019), they examined whether it was possible to reproduce participant-provided duration estimates from accumulation of relatively large changes in the activity of sensory cortices of human participants, rather than in the proxy neural-network-based model previously used. They found that biases in human duration judgments could indeed be predicted by accumulation of these changes in BOLD activity in participants' visual cortex (but not in control models based on auditory or somatosensory cortex). This result supported the previous finding but in a biological network, and showed that a correspondence between the content being experienced by participants and the cortical region being examined was key to reproducing participants' subjective duration judgements.

Finally, Fountas and colleagues (2022) combined and extended perspectives from the episodic memory literature and these models of human time perception by developing a novel hierarchical Bayesian network based on predictive processing principles. This model of episodic memory and time perception was based on the similar premise that relatively large changes in neural activity can be used as the basis for time perception, but took this a step further to explicitly implement a model of memory - including event segmentation - incorporating episodic and semantic memory based on the segmented event boundaries. By doing so, Fountas et al. (2022) were able to replicate features of human time estimates that depend specifically on memory, as often described in studies of retrospective time perception (e.g. Block & Zakay, 1997). Compared with the data of ~13000 human participants providing prospective (knowing they were estimating duration) or retrospective (not knowing until after that they would need to estimate duration) estimates, the model could reproduce many of the biases in human estimates, including those related to video scene (whether a busy or quiet scene), cognitive load (whether participants were required to estimate only duration or needed to also attend to another feature of the scenes), and whether they estimated duration based in the moment or in retrospect. This study provided an overarching method through which to integrate the fields of time perception and episodic memory, within a hierarchical predictive coding framework, with event boundaries at the core of understanding both topics. Throughout this series of papers, it was noted that the measure of change between successive moments - termed salient changes or events - conceptually, or in the case of Fountas et al. (2022) explicitly, corresponds to moments of surprise in an information theoretic sense. Consequently, what is termed "salient event" in that research thread maps onto "event boundaries" as suggested in the event segmentation literature. This interpretation is supported by recent work (Kumar et al., 2022) demonstrating that event boundaries are associated with transient increases in Bayesian surprise in a text generation neural network (GPT-2) while input an audio-based narrative. Therefore, in this paper we use the term "event boundary" or event annotation when referring to the segmentation sequences/events obtained, consistent with previous studies in the episodic memory literature (see Fountas et al., 2022, for further discussion of this correspondence between approaches).

To reconcile the described research threads, the present study compares event boundaries obtained by three different methods (see also Figure 1):

> 1) as detected by data-driven methods in neuroimaging (EEG) data, obtained while human participants watched naturalistic videos (Geerligs et al., 2021);
> 2) as detected by a deep neural-network-based model, previously developed in the context of modelling human time perception (Roseboom et al., 2019; Fountas et al., 2022) and applied by directly processing the video clips;
> and 3) as provided by human participants in an online experiment.

We compare event segmentations provided by the data-driven and model-based methods with human annotations to understand their ability to reproduce human-provided segmentation sequences, and further, to contrast performance on this ability between approaches. Importantly, all of this will be done using naturalistic videos (e.g. as in Roseboom et al., 2019), rather than TV/movies, to avoid potential issues with highly edited content (e.g., Magliano & Zacks, 2011 and see Cutting, 2014 for a discussion about the complex relationship between editing techniques and event segmentation).

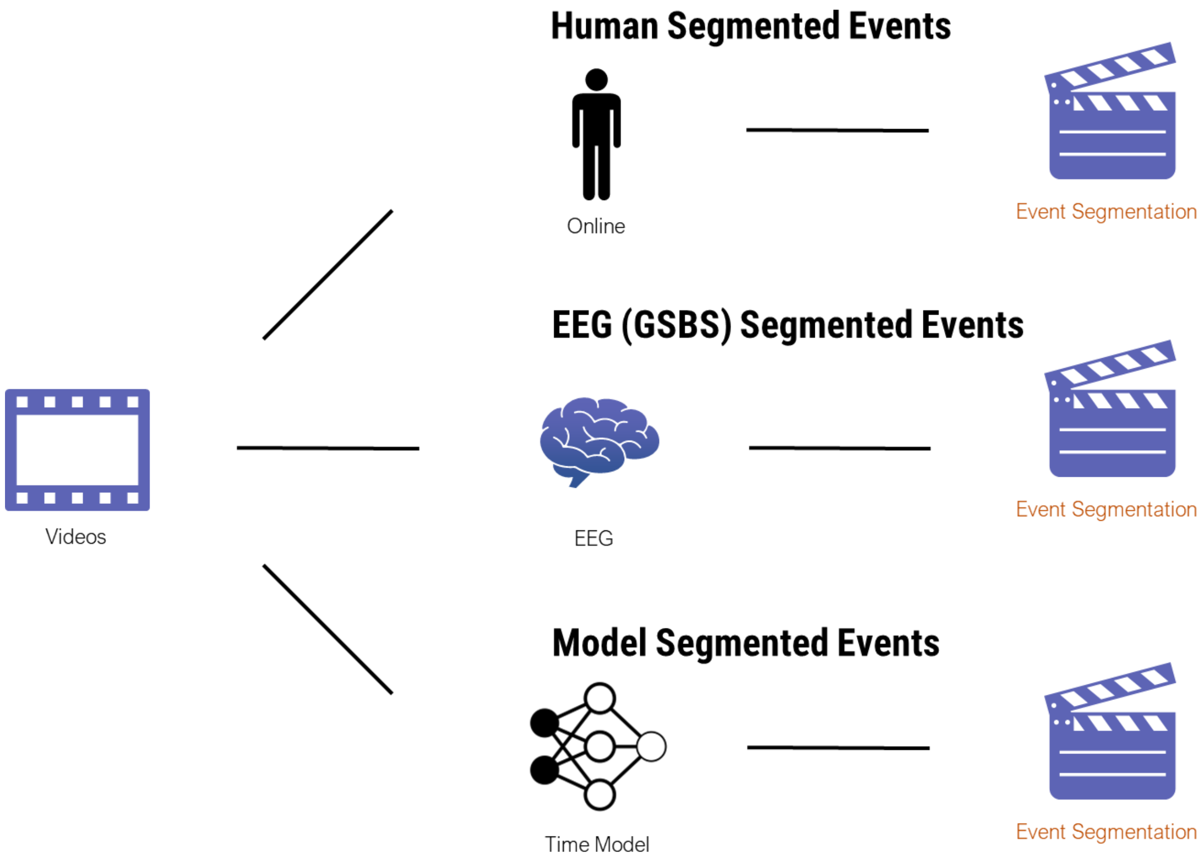


Figure 1. A schematic of the study design. The study consists of three methods of event segmentation (Human, GSBS, model-based), all conducted on the same, naturalistic videos (up to 64 seconds long), and obtained through two different experiments. In an online study, we had human participants watching silent video clips and inserting event boundaries when they believed something had occurred in the video. In a separate in-lab EEG experiment, we had human participants watching these same clips and subsequently, using the data-driven GSBS segmentation algorithm, we processed the EEG data to obtain event segmentations. Our third event segmentation method was based on a model developed to predict human duration judgements. This model-based method provides event annotations based directly on the frames of the video clips.

Methods

EEG Experiment

Participants

Fifteen adults were recruited to participate in the study, which was approved by the University of Sussex ethics committee (Sciences & Technology C-REC - reference code: ER/WJR22/9). Participants were recruited via the SONA system at University of Sussex and were compensated £12 or 8 credits for their time. Due to technical issues, 2 participants had to be excluded from the analyses. Specifically, one participant only completed 60 trials instead of 80 (3 blocks instead of 4) and the markers of EEG and eye tracking data structures of the second participant were misaligned.

Stimuli

Stimuli were based on videos obtained in the city of Brighton (United Kingdom), the University of Sussex campus, and the local countryside. The video segments were recorded with a GoPro Hero 4 camera at 60 Hz and 1920x1080 pixels from face height, and further processed at 30 Hz and 1280x720 pixels. The brightness/contrast of the videos was not controlled. Clips employed in the experiment were extracted from a pseudo-random list of 4290 trials generated out of the total cumulative video duration (165 minutes - 33 videos of 5 minutes each), which was structured into 330 repetitions of 13 durations ranging from 1 to 64 seconds (1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64 s). This means that for each trial, one video and one trial duration were uniformly-randomly selected and the equivalent number of frames was also uniformly-randomly assigned to frame locations within the respective video. While each participant had broadly the same stimuli presented in terms of durations seen and distribution of scene types, the stimuli shown were not identical - each participant had a different, though potentially overlapping stimulus set. This point becomes important to keep in mind when considering the analysis methods applied in further sections.

The videos could be broadly classified into three video types in terms of content: videos recorded while walking around the city, scenes recorded while walking around the campus and outside in the campus green zones, and quiet scenes in an office or café (Figure 2). This classification is based on previous results for duration judgements with these same videos (Roseboom et al., 2019; Suárez-Pinilla et al., 2019) and each video type generally has a greater amount of perceptual change than the next (city>campus and countryside>office or café). Perceptual change can be defined at an abstract level as the amount of change between consecutive video frames – more complex and dynamic videos will have greater perceptual change than static scenes involving inanimate objects. The computational basis of ‘perceptual change’ was previously established by Roseboom et al., (2019), as described in the Introduction. Specifically, duration estimates produced by the model (which were a transformation of the accumulated perceptual change, converted from arbitrary ‘change units’ into seconds by support vector regression) deviated from the mean by +24%, –4% and –7% for city, campus/countryside and office/café videos, respectively.

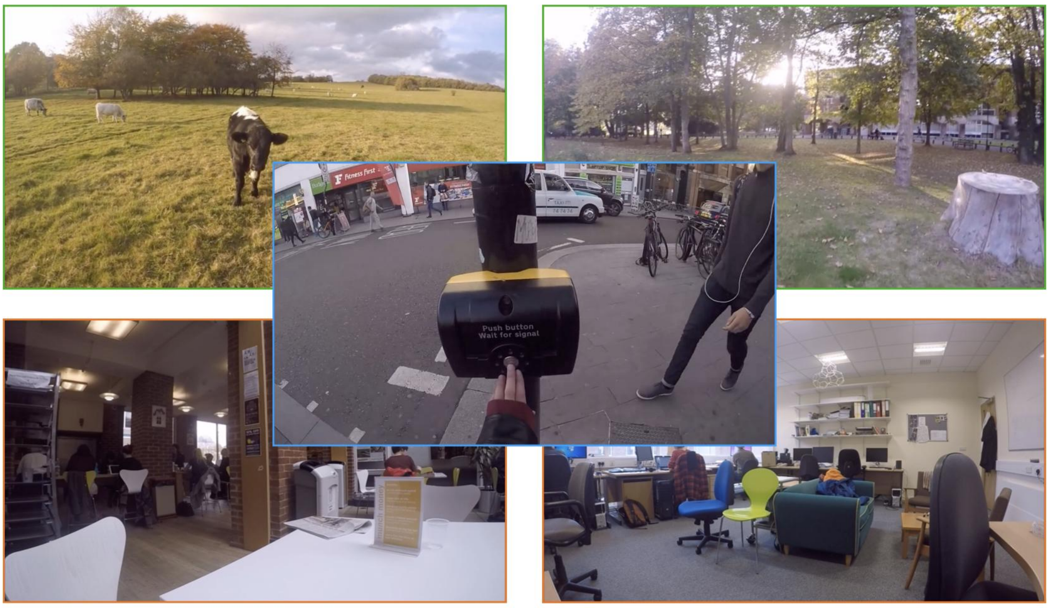


Figure 2. Experimental Stimuli. The figure shows thumbnails of the 5 possible scene types used in the experiment, grouped in the 3 main macro categories of scene content by means of a coloured frame. Central panel: city (blue). Upper panels: countryside and campus (green). Lower panels: café and office (orange).

Design and Procedure

Participants were instructed to watch video segments of natural (campus/countryside landscape), urban (city) and interior (cafe´ and office) scenes, while their gaze patterns and electrophysiological signals were recorded. Each participant completed 80 pseudo-randomized trials (apart from a single participant who only completed 60 trials due to technical issues) in four blocks of 20 trials. Each block took approximately 12 min to complete, with the entire session lasting approximately 1 hour. Clip duration and content (type: city, campus/countryside, office/café) were not perfectly counterbalanced, but each participant viewed at least one clip per duration.

Each trial started with participants pressing the left mouse key to begin presentation of the video stimulus. Once the clip was completed, participants were requested to evaluate its duration by moving a cursor over a visual analogue scale ranging between 0 and 90 s (linearly spaced along the scale in steps of 1 second with a resolution of 0.1 seconds). Responses were confirmed by pressing the left mouse button. Participants were instructed to not explicitly count during the stimulus presentations. They were told that counting included physical rhythmic tapping, or the mental equivalent. At the end of the trial participants rated their engagement level on a Likert scale ranging between 1 and 5. This subjective engagement data is not used in the presented analyses.

Data Acquisition

The experiment was created using Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in MATLAB 2012b and the Eyelink Toolbox (Cornelissen et al., 2002). The presented stimuli were displayed on a LaCie Electron 22 BLUE II 22” (1280 x 1024 pixels screen resolution) at a 60 Hz refresh rate. Electroencephalography (EEG) was acquired at a sampling rate of 1024 Hz with an ANT Neuro amplifier system (ANT Neuro, Enschede, Netherlands). EEG activity was measured continuously from 64 active electrode Ag/AgCl electrodes arranged according to the International 10-5 system over the scalp (Waveguard cap, ANT Neuro). The ground electrode was placed on participants’ forehead and data were averaged across the whole head online. Impedances were kept below 10 kΩ throughout the experiment session. Blood-volume pulse (BVP) measurements were obtained using a BVP-Flex/Pro (9308M) sensorand FlexComp System (T7550M) from ThoughtTechnology (Montreal, Quebec, Canada). Eye tracking was concurrently acquired on participants’ dominant eye with an Eyelink 1000 Plus (SR Research, Mississauga, Ontario, Canada) at a sampling rate of 1000 Hz. Calibration of the eye-tracking system was performed at the beginning of each 20-trial block, with a standard 9-point grid and a maximal average error of 0.5 degrees of visual angle (dva).

Participants sat in an electrically shielded faraday cage for the entire experiment and their head was stabilised with a chin and forehead rest at 57 cm distance from the stimulus screen.

Pulse-oximeter and eye tracking data won’t be reported on in detail in this paper: eye-tracking data have solely been used to optimise EEG preprocessing. See Suárez-Pinilla et al. (2019) for analysis of larger dataset using similar videos and looking at potential relationships between features of heart rate and eye behaviour (saccades and pupil) and time perception.

EEG-Eye tracking Synchronisation

We synchronised our EEG and eye-tracking data on a block-by-block (session-by-session) basis by means of custom MATLAB (R2019a) code and the EYE-EEG toolbox (Dimigen et al., 2011). Synchronisation between the EEG and eye-tracking is not only important to analyse evoked potentials generated in response to saccades and fixations, but it's also necessary to ensure proper cleaning of the neurophysiological data. As reported by Dimigen (2019), this is particularly important for experiments that involve the use of naturalistic stimuli such as videos and movies since they're characterised by an increased number of eye movements that could cause distortions in the EEG data. While eye movements are also present during classical experimental settings, these artifacts are particularly problematic for the aforementioned naturalistic cases and they can generally be divided into 3 classes. First, corneoretinal dipole (CR) generated by the difference in voltage between the front (cornea) and back (retina) of each eyeball that linearly increases with saccade size (Marmor & Zrenner, 1993). Second, eyelids related artifacts generated by the lag during blinks and/or after oblique-upward saccades (e.g., "rider artifact" - Lins et al., 1993; Matsuo et al., 1975) and resulting in a brief frontal positivity (Plöchl et al., 2012). Finally, saccadic spike potentials (SP) which are very brief high-frequency biphasic waves ramping up approximately 5-10 ms before each saccade (Blinn, 1955; Keren et al., 2010). SPs are thought to be caused by myogenic (EMG) activity of the motor units of the extraocular muscles at saccade onset and represent the most problematic artifact to correct (Boylan & Doig, 1989).

EEG Preprocessing

The EEGLAB toolbox (version 14.1.0.b) running on MATLAB (R2019a) was used to preprocess the data (Delorme & Makeig, 2004) on a block-by-block basis (approximately 12 minutes of data per block - coherently with the synchronisation with the eye tracker). Data were first downsampled to 512 Hz and were then filtered between 0.1 and 40 Hz (FIR filter; -6db response at 50% of frequency) to eliminate slow drifts and the remaining amount of 50 Hz line-noise caused by the equipment inside the Faraday Cage.

Bad channels were removed following three main criteria based on the FASTER approach: high variance, low correlation with other channels and abnormal Hurst exponent (Nolan et al., 2010). Specifically, channels were z-scored and rejected if they exceeded a threshold of 3 for each of the aforementioned criteria. Noisy channels were then spherically interpolated. Subsequently, data were re-referenced to the average of all channels (excluding EOG channels). To ensure perfect correspondence between the timeline of each clip and the respective neurophysiological activity, no EEG segments related to video watching were removed. Subsequent cleaning procedures were based upon artifactual components removal by means of Independent Component Analysis (ICA).

Previous studies showed that training ICA on more than 1 Hz high-pass filtered data improved weights estimation and consequent extraction of related components (Winkler et al., 2015). Because of this reason, ICA optimised for removal of oculomotor artifacts (OPTICAT - Dimigen, 2020) was trained on a copy of the original dataset that was filtered between 2 and 40 Hz (FIR filter; -6db response at 50% of frequency - same rejected and interpolated channels as the original dataset) and resulting weights were applied to the original - more conservatively filtered - dataset. In addition, as suggested by Dimigen (2020), we overweighted EEG samples related to saccades and fixation on the training data to optimise extraction of ocular components. Artifactual components related to blinks, saccades and general ocular artifacts (see above) were then removed by means of the *pop_eyetrackerica* function (EYE-EEG Toolbox - Dimigen et al., 2011) which automatically ranks bad components based on the variance ratio between components related to saccades and fixations (threshold: 1.1 - Plöchl et al., 2012).

Finally, consistent with previous studies (Silva et al., 2019), before the event segmentation analysis, cleaned data were downsampled to 64 Hz to reduce computation time and z-scored such that the mean of each electrode was equal to zero.

EEG Analysis - Identification of Event Boundaries

Event boundaries in neurophysiological data were identified with the Greedy State Boundary Search (GSBS) algorithm developed by Geerligs and colleagues (2021). The algorithm identifies the location of state boundaries iteratively by leveraging the correlation between the activity in each timepoint and the mean activity of all timepoints in its corresponding estimated state. A state boundary is then fixed in correspondence with the highest fit (correlation value). In addition, to identify the optimal number of events (k) the authors created a metric called t-distance which reflects the distance between the distribution of the within-state timepoints' correlation and the between *consecutive* states one (using t-statistic). The purpose is to maximise the similarity (correlation) between timepoints in the same state while minimising similarity of timepoints across consecutive states.

GSBS is not designed to identify recurrent states (the first time point is always within event 1 and the last one with event k; Geerligs et al., 2021) and for our analysis we set the value of *k* equal to the number of seconds per clip representing the upper bound of the model (segmentation could find stability at lower numbers of k). We fit GSBS separately on the z-scored EEG data of participants associated with the presentation of each clip. We then examined the distribution of t-distances across the potential number of states (events) and selected the number of events associated with the maximum t-distance value. The precise timing in seconds of each event boundary was then obtained by dividing its corresponding value in data points by the EEG sampling frequency (64 Hz at this stage of the analysis). While the value of t distance (and the associated number of events) would be better estimated after averaging across participants and/or within a cross-validation procedure (Geerligs et al., 2021), the fact that each participant watched a unique set of naturalistic videos without repetitions rendered these procedures unfeasible.

Our event segmentation analysis followed a similar approach to the one presented on the Naturalistic Data Analysis portal (https://naturalistic-data.org/content/Event_Segmentation.html), but it featured different metrics and procedures to compare model with human annotations (see "Comparing Event Boundaries between Modalities"). The entire analysis was run in Python (v. 3.8.5) via custom code and the *statesegmentation* package (v.0.0.5 - https://pypi.org/project/statesegmentation/; Geerligs et al., 2022; Geerligs et al., 2021).

Computational Modelling

Model Architecture and Procedure

Our computational model is composed of the main components of the model reported in Roseboom et al., (2019): i) an image classification deep neural network (DNN), ii) a network state comparison stage, iii) a threshold mechanism and iv) a set of accumulators.

We deployed a pre-trained version of the convolutional neural network MobileNetV2 (Sandler et al., 2019) that is available in Keras Applications (Chollet & others, 2015). The network was trained to classify the images of the LSVRC-2010 ImageNet dataset into 1000 different classes. The network is composed of one initial convolutional layer followed by 19 residual bottleneck layers while the final layer outputs a probability for each of the classes of stimuli processed during training.

Following Fountas et al (2022), the model used here is viewed as a hidden Markov model and the image classification network is treated as a pre-optimized amortised approximate posterior for the latent state of the environment. Under this interpretation, the prior model of this state at time *t* is equal to the state posterior at time *t-1*, practically assuming that no change has occurred. Therefore, the Kullback-Leibler (KL) divergence between successive network states can be used to quantify the level of surprise (or belief update) across time (Roseboom and colleagues (2019) used Euclidean distance as measure of change).

We fed the DNN with the same clips in the same order in which these were seen by participants completing the original EEG experiment. A single frame of every clip was processed at each timestep (30 Hz) by the network thus resulting in a series of activations across layers and in a final vector of class' probabilities in the output layer. By operationalising the output of the DNN as an approximate posterior categorical distribution with 1000 dimensions, the difference between successive frames/steps was computed by calculating the KL divergence between posterior and prior at each timestep (respectively the output probabilities for frames at t and t-1, quantifying the belief update or surprise at timestep t). Salient perceptual events were detected and accumulated every time the KL divergence between successive timesteps crossed a stochastically decaying threshold. The threshold was reset to its maximum value every time the KL divergence value exceeded it and it was computed by means of the same formula used by Roseboom et al. (2019):

$$T^k_{t+1} = T^k_t - \left(\frac{T^k_{\max} - T^k_{min}}{\tau^k}\right) e^{-\left(\frac{D}{\tau^k}\right)} + N\left(0, \frac{T^k_{\max} - T^k_{min}}{\alpha}\right) \qquad (1)$$

where T^k_t is the threshold value of the kth layer at timestep t, and D represents the number of timesteps since the last time the threshold value was reset. T^k_{max} and T^k_{min} are the maximum and minimum threshold value, respectively, and τ^k the decay time constant for the kth layer. Stochastic noise drawn from a Gaussian was added to the threshold and α — a dividing constant to adjust the variance of the noise. While Roseboom and colleagues (2019) extracted activation for four different layers (see Table 1 in Roseboom et al., 2019) by means of a different specification of T_{min} and T_{max} per layer, here we solely focus on the output layer of the network for two main reasons. First, because we only have one set of event boundaries given by each of the other methods (Human annotations and GSBS segmentations) with which to compare. Second, because for the output layer (and not the other layers) it is straightforward to interpret and treat states as probability distributions (e.g., since the output is composed of 1000 logits that can be converted to a categorical distribution via normalisation).

To make sure that our results were not strictly contingent/dependent on the specific values of the threshold parameters, we reran the thresholding algorithm by sampling a space of possible values of both T_{min}(ranging from 1 to 5) and T_{max}(ranging from 6 to 10) in steps of 1. The selected ranges for both T_{min} and T_{max} were set to cover the minimum and maximum KL divergence values extracted by the output layer of the network across all clips. Performance of the model across the possible combinations of parameters was then assessed by the magnitude of correlation between the EEG participants' reported duration of each clip (in seconds) and the duration estimated by the model (see Supplementary Material, Figure S1 for details). To emphasise, this means that the model parameters were chosen to optimise correspondence with the duration estimates provided by human participants in the EEG experiment, not the event annotations provided by human participants (see next section). The model that was selected as the main one reported in this manuscript featured $T_{min} = 1$ and $T_{max} = 10$.

Online Behavioural Experiment

Participants

Two hundred twenty-four adults (18-35 years, gender unknown due to anonymization procedure) were recruited online via Prolific (https://www.prolific.co/) to participate in the study, which was approved by the University of Sussex ethics committee (Sciences & Technology C-REC - reference codes: ER/AM2049/5 and ER/AM2049/7). Participants were compensated £9/h for approximately an hour of their time. No attempt was made to prevent the 15 participants from the EEG experiment from participating. However, the experiments were run more than 2 years apart, with one run on campus and the other on Prolific and no attempt to recruit or notify the original 15 participants was made. After having applied our exclusion criteria (see Exclusion Criteria below for details), we retained 154 participants. Of those, only 131 were included in the presented analyses, corresponding to the number of raters matching the 13 used stimulus sets (EEG Experiment - Participants).

Stimuli

The stimuli were the same as used in the in-lab EEG experiment. This included the fact that there were 15 overlapping, though not identical stimulus sets (see EEG Experiment - Stimuli). Therefore, for these participants, there were effectively 15 (13 after exclusion) subsets of data obtained.

Design and Procedure

The study was a simple online behavioural experiment. Participants were requested to watch the same series of video clips depicting natural scenes (e.g. walking around a city, walking in the countryside/campus, or sitting in a quiet office or a café) that were used in the aforementioned EEG experiment (see above). Their task was to place event boundaries whenever they thought a strong perceptual or semantic change took place in the scene. More precisely, participants instructions were the following:

"Welcome! In this experiment, you will watch a series of videos. Your task will be to segment each video into a series of events. Events can be defined as meaningful sub-sections of each clip. Events are determined by moments of perceptual and/or semantic change in the scene.
For instance, an event boundary can be inserted whenever something captures your attention: a child running past, a car suddenly entering view or a person exiting the field of view etc..
Click the "Next" button for more instructions.

Segmentation will be possible by means of the "Add Event" button placed below the video.
All you have to do is to click it when you think that a meaningful change has happened in the video. In that moment the button will turn red to confirm that the event has been inserted.
At the end of the clip, you will be given the possibility to revise your segmentation by means of a continuous timeline that you can use to play/stop specific section of each clip, just like you have seen on the training video.

New Events can be added by means of the same "Add Event" button and incorrect ones can be removed as you please by clicking on the "Remove Event" button while the timeline's cursor is on the previously placed event.

Some clips will be extremely dynamic whereas others could be very short/still and won't contain very much at all. You might not feel that there was any change to note. This is fine.
Click the "Next" button for more instructions.

Please note that all data will be subject to quality control checks at the end of the session, in order to check for excessive missing events and random/repeated responses (e.g. including a random number or the same number of events repeatedly irrespective of the content of the video). Payment will only be issued if all trials are fully completed according to instructions.
Click the "Next" button to start the experiment"

As mentioned in the instructions, participants were able to place events while watching each clip just by clicking on the "Add Event" button on screen below the clip. In addition, events could be increased, moved and removed retrospectively as soon as the clip ended by means of a dynamic timeline placed below it. For the primary analyses reported here, we use the final set of events provided by participants, including any self-edits made after initial segmentation. An example of the experiment structure can be seen here: https://www.youtube.com/watch?v=qxWORNayxJk.
The experiment was created with JsPsych (v6.03 - de Leeuw, 2015) and custom Javascript code and it was hosted on Cognition (https://www.cognition.run/) before being deployed on Prolific.

Exclusion Criteria

To maximise data quality and avoid the presence of online bots, we only recruited participants from a selected pool having an approval rate of more than 90% for previous experiments on Prolific. In addition, while event segmentation tasks don't typically have an objective score and/or performance to optimise, participants had to be engaged and follow specific guidelines (see "Design and Procedure").
To detect when participants left trials containing substantial events empty and/or when they randomly inserted events regardless of stimulus content, the number of estimated events (continuous) was estimated by means of a linear regression separately fitted to each of them (80 trials). Models were fitted iteratively in R via the *tidymodels* package (Kuhn & Wickham, 2020 - with lm from the *stats* package as engine). Video duration (13 possible durations, operationalised as continuous in the model) and scene label/type (categorical - 3 levels:

1:city, 2:campus/countryside, 3:office/café) were the two independent variables/predictors. Difference between mean number of events per level was assessed by means of a contrast analysis (e.g. *estimate_contrasts()* function - *modelbased* package; Makowski et al., 2020).
The analysis of the online behavioural results and corresponding exclusion criteria were pre-registered on OSF (osf.io/pr49b).
Participants passed exclusion criteria if and only if:

1. There was a significant main effect of clip duration on the number of event boundaries reported by participants. Specifically, longer scenes should contain more events irrespective of content (positive beta/slope).
2. There was a significant main effect of scene type on the number of event boundaries reported by participants. We expected clips filmed around the city to include more event boundaries (than office videos) since they include a lot of dynamic transitions and interactions with people and objects. Specifically, this should have been reflected in a significant mean difference of the estimated number of events between these two levels (respectively 1 and 3) after post hoc tests.
3. The number of trials containing no event boundaries didn't exceed 70% of the total amount.

In total, based on these criteria we excluded a total number of 70 participants (70/224 - 31.25 %).

Data Analysis

Comparing Event Boundaries between Modalities

The main question that we wanted to tackle with this study was whether the event boundaries extracted by means of the two methods that we deployed - GSBS on EEG data and model-based analysis of videos - corresponded to the ones provided by human participants. Therefore, we needed to devise a method by which to discriminate when the annotations produced by these approaches could, statistically, be considered to be similar or dissimilar to human raters. Specifically, we wanted to answer the question of whether, in general and across the stimuli that we used, the GSBS or model-based event segmentations could be discriminated from a group of human raters segmenting the same videos. In order to test this, we developed two statistical approaches, with the aim of determining whether the GSBS or model-based annotations were statistically distinct from a group of human raters.

It is important to note that in both approaches the null hypothesis is that segmentations generated by either method resemble segmentations generated by human participants when segmenting the same videos, while the alternative hypothesis is that they did not. This entails determining a means to quantify the "distance" between segmentations. Of note, the two distance measures that we adopted are completely orthogonal to the two inferential approaches we developed (i.e., either metric could have been used with either statistical approach).

Direct Approach: Mean-squared Distance by Scene Type and Stimulus Set

The main idea behind our first approach was to operationalize the event segmentation process as a point-process (with boundaries as points in times *t1* < *t2* < ... < *tn*). The associated counting process is the process *N(t)* in continuous time defined as:

$$N(t) = \max\{k : t_k < t\} \tag{2}$$

N(t) counts the number of events which have occurred before time *t*. Plotting *N(t)* against t gives rise to a graph that looks like a series of uneven steps (see Figure 3).

Given two point processes, a possible measure of distance between them is simply the area (mean-squared distance) between the corresponding counting processes:

$$d = \sqrt{\int_{t=t_{start}}^{t_{end}} [y(t) - x(t)]^2 dt} \tag{3}$$

where *y* and *x* represent the two counting processes and *t* corresponds to time. Note that this metric is not necessarily time-symmetric: if the segmentations have different numbers of event boundaries, then if both segmentations are time-reversed the mean-squared distance will not generally be the same.

An alternative approach would be to match events according to a prespecified time threshold, and then count non-matching events and normalise by the total number of events. We rejected this approach because firstly it introduces a somewhat arbitrary parameter (the time threshold), and secondly it essentially neglects inter-event durations. In principle we could devise a permutation test for the null hypothesis of zero segmentation distance, by permuting event boundaries whilst maintaining inter-event durations. However, we considered this unviable due to the large number of relatively short clips, which risk a large number of repeated boundaries.

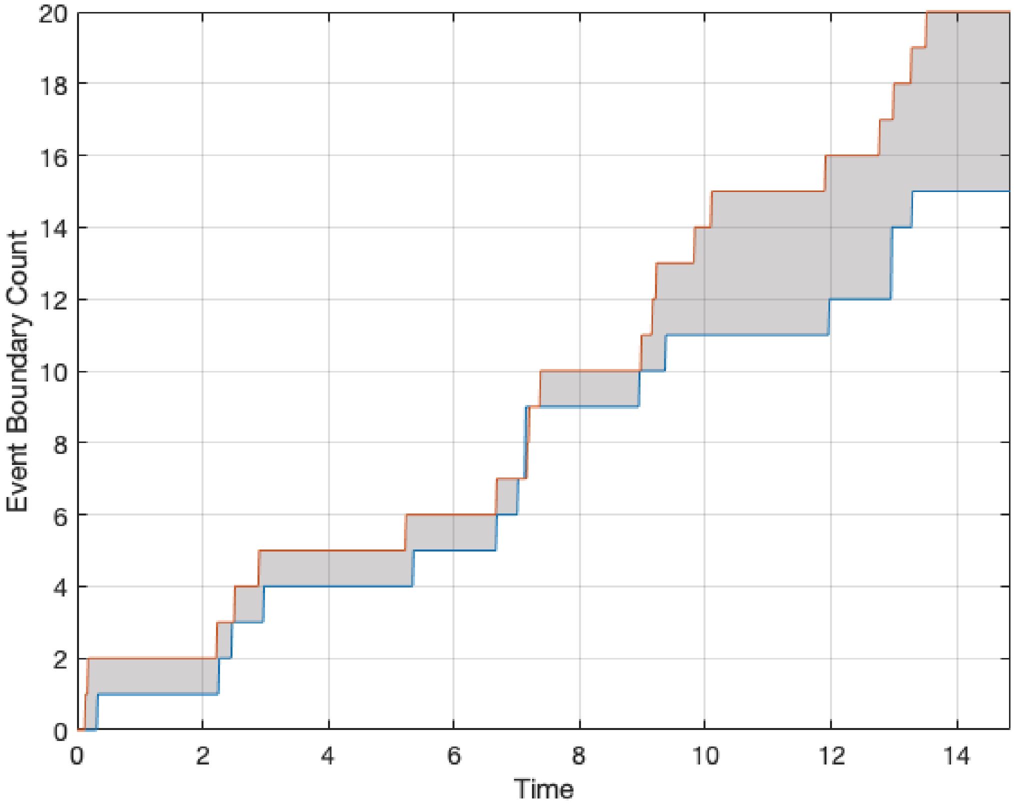


Figure 3. Event Segmentation sequences of two simulated raters. The orange line indicates the segmentation sequence of one rater; the blue line another, treating both as a counting process (*N(t)*) over Time (*t*), such that as time continues (x-axis), the number of events counted increases (y-axis). The area between them in grey is the key measurement in the Direct Approach.

We thought that a more useful way to cast this problem was to compute the distance between the segmentation process of each human rater, between EEG (segmentation completed via GSBS) and human raters, and between model-based and human raters. The distribution of human inter-rater distances would then represent our null distribution against which to compare the distances of each of the other two methods relative to human raters. This allows us to obtain an inference (per clip) regarding whether the non-human event segmentation methods would appear as outliers compared to the distribution of human raters. A segmentation generated by a given method (GSBS or model-based) for one specific clip would be considered to be significantly different from human annotations if and only if the average distance value between this method and all human segmentations was larger (e.g., more extreme) than 95% of the inter-rater human distances distribution (α=0.05).

Given the restricted amount of data (10 online raters and a single EEG/model-based segmentation per unique clip) p-values were computed by aggregating inter-rater distances by scene category (city, countryside/campus, office/café) thus resulting in 3 p-values per stimulus set/participant. This was advantageous for two reasons: first, it allowed us to build more reliable and representative null-distribution (given the 3-fold increase in inter-rater distance values); second, it gave us the opportunity to test the hypothesis that event segmentation is differentially modulated by the perceptual content across scene categories. Crucially, for the analysis we only focused on clips longer than 10 seconds (in our specific case between 12 and 64 seconds). We decided to only retain clips longer than 10 seconds i) to avoid the possibility of including clips that don't contain any event boundaries at all and most importantly ii) because very short clips might be too hard to segment for both human participants and models thereby skewing our analysis towards extreme cases (this selection was decided before we collected data for the online experiment; see our preregistration: https://osf.io/4g785).

In order to rule out the possibility that our different segmentation methods were not significantly different from the distribution of human raters because all raters (human, GSBS and model-based) were simply producing a random series of annotations, we ran a control analysis by generating a series of a 100 random raters, each with their own segmentation profile. For each simulated rater, we first sampled a discrete random number of event boundaries between 0 and the current clip duration. Then, we randomly sampled the timing of each boundary from a continuous uniform distribution that also ranged between 0 and the current clip duration. Importantly, since it's very unlikely that humans would generate multiple meaningful events within a very short period of time, we constrained our model such that event boundaries had to be separated by at least 1 second. Consistent with our original analysis, we then computed a single inter-rater-distance value between random raters and the original human raters per clip (thus obtaining 1000 inter-rater distance values). Finally, we averaged all these values within scene-type and compared them with the null distribution built upon the inter-rater distances of the human raters to obtain our p-values.

If the number of instances where the random event annotation process could be discriminated from our human rater distances was lower or similar to those obtained for our EEG data and model-based approaches, then we would conclude that our method of comparing rater distances is insensitive and no conclusions could be drawn. However, if the random generated annotations could be discriminated from our human rater distances in the majority of cases, then this would suggest that our method of determining whether a given set of distances (e.g., a model or an ensemble of random generators) is outside of what we would expect for our human raters is working.

Meta Approach: Average Minimum Distance and Mann-Whitney test

The rationale behind our next test is as follows: ideally, to establish to what extent the GSBS (resp. model-based) segmentations are "human-like", we would like to test the null hypothesis that the GSBS (resp. model-based) segmentation for a given video clip is drawn from the same distribution as human segmentations for the same clip. This essentially requires an explicit segmentation model, and appears to be highly non-trivial, in particular because the set of all potential segmentations over a given time interval is somewhat awkward to deal with analytically, and choice of a suitable class of segmentation models is not obvious. While we are investigating this approach for a future study, here we choose a more tractable null hypothesis: that the distance between the GSBS (resp. model-based) segmentation for a given video clip and human segmentations of the same clip, is drawn from the same distribution as the distance between two arbitrary human segmentations of that clip. Note, crucially, that if we reject this alternative null hypothesis then logically we must reject the ideal null hypothesis. The converse, however, is not true; if we fail to reject the alternative null, it does not follow that we would necessarily have failed to reject the ideal null. Testing the alternative null hypothesis thus has reduced statistical power as compared to testing the ideal null hypothesis.

There are several other differences between our implementation of the above Direct approach and the Meta approach we describe below. First, here we used the average minimum distance instead of the mean-squared distance between events as the (dis)similarity metric. The measure is defined as follows: let $t_i^{(1)},\ i = 1,..., n_1$and $t_j^{(2)},\ j = 1,..., n_2$ be the timestamps of the event boundaries in segmentation 1 and segmentation 2 respectively[1]. Then the distance between segmentations is

$$d = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} \min_{j=1}^{n_2}\left\{\left|t_i^{(1)} - t_j^{(2)}\right|\right\} + \sum_{j=1}^{n_2} \min_{i=1}^{n_1}\left\{\left|t_j^{(2)} - t_i^{(1)}\right|\right\}\right] \tag{4}$$

Intuitively, for each event in segmentation 1 we find the nearest-in-time event in segmentation 2, and store the time difference. This is repeated with segmentations swapped, and all minimum time differences are then averaged. A principal difference between this and the mean-square distance metric introduced earlier, is the manner in which disparities between *number* of events in the two segmentations is penalised. Note that unlike the means-square distance, this measure is invariant under time reversal. The second difference between approaches was that null distributions were based on surrogate raters generated using Poisson processes. Finally, we assessed statistical significance by means of Mann-Whitney tests and the computation of p-values on the distribution of their scores rather than on the distribution of distances per se (hence we refer to this method as “Meta” rather than “Direct”). The implementation of the method follows two parallel streams: computation of a single Mann-Whitney test statistic based on the experimental data and generation of a surrogate null distribution of Mann-Whitney scores based on randomly generated data.
The procedure is as follows: for each video clip

a. Calculate the average minimum distance from the GSBS (resp. model-based approach) segmentation to each of the 10 human segmentations of the clip. Call this set of distances D1.
b. Calculate the (10×9)/2 = 45 distances between human segmenter pairs. Call this set of distances D2.
c. Calculate the Mann-Whitney statistic for stochastic dominance D1 > D2, and store as *mwstat* (a single number).

Following our main question, we would like to know how this Mann-Whitney statistic would be distributed under the alternative null hypothesis described above. However, there are two

[1] We add dummy “event boundaries” at the start and end times of the video clips.

problems with this: i) While an analytic asymptotic distribution of the Mann-Whitney sample statistic is known, the sample sizes (10 vs 45) are likely not big enough to be reliable (Siegel, 1956); also, more seriously (ii) the samples violate an important assumption behind the null distribution: specifically, the 10 model-human distances are not mutually independent, and nor are the 45 human-human distances; and nor are the model-human and human-human distances independent of each other. In fact all distances are constrained by the distance-metric geometry[2]. We thus generate a surrogate null distribution.

Construction of the Meta surrogate null distribution attempts to mimic the empirical distribution of human-human segmentation distances as closely as possible. For each clip, for sample s of $S = 10000$, we perform the following:

1. Pick one of the 10 human segmentations for the given clip uniformly at random. Generate a random segmentation as a Poisson process with expected number of event boundaries equal to the number of event boundaries in the selected human segmentation.
2. Repeat step 1 independently 10 times.
3. Calculate and store the 10 distances between the segmentation generated in step 1 and each of the segmentations generated in step 2. Call this set of distances D01.
4. Calculate and store the (10×9)/2 = 45 distances between all pairs of segmentations generated in step 2. Call this set of distances D02.
5. Calculate the Mann-Whitney statistics for stochastic dominance D01 > D02, and store as *mwstat0(s)*.

We now have a surrogate null distribution of Mann-Whitney distance-comparison statistics *mwstat0*, which is “human-like” in the following sense: although the Poisson segmentations pay no attention to the actual clip, the distribution of number of identified event boundaries is approximately the same as the empirical distribution of number of event boundaries identified by human segmenters for the given clip. As such, the distribution *mwstat0* is a reasonable surrogate for the sample distribution of Mann-Whitney stochastic dominance statistics for human-human segmentation distances.

[2] While it is unclear whether either of our distance measures is a metric in the strict mathematical sense, they nonetheless impose constraints on inter-segmentation distances among groups of segmentations.

We now calculate the p-value for *mwstat* (steps a-c above) in the surrogate null distribution *mwstat0*. Rejection of the null hypothesis based on this p-value (i.e., *mwstat* is an outlier in the *mwstat0* distribution) is interpreted as "The GSBS (resp. model-based approach) does not behave like a human segmenter for the given clip". Examples of the outcome of this process can be seen in Figure 4, for both a case wherein the null–hypothesis would be rejected, and one where it would not be.

As there are multiple (i.e, per-clip) tests, statistical inference must be corrected for multiple (in our case independent) hypotheses; we used Bonferroni correction. Note that, as compared to uncorrected inference, multiple-hypothesis correction, perhaps counter-intuitively, makes it less likely that the “human-like” null hypothesis will be rejected (see Supplementary Materials, Tables S6 and S7).

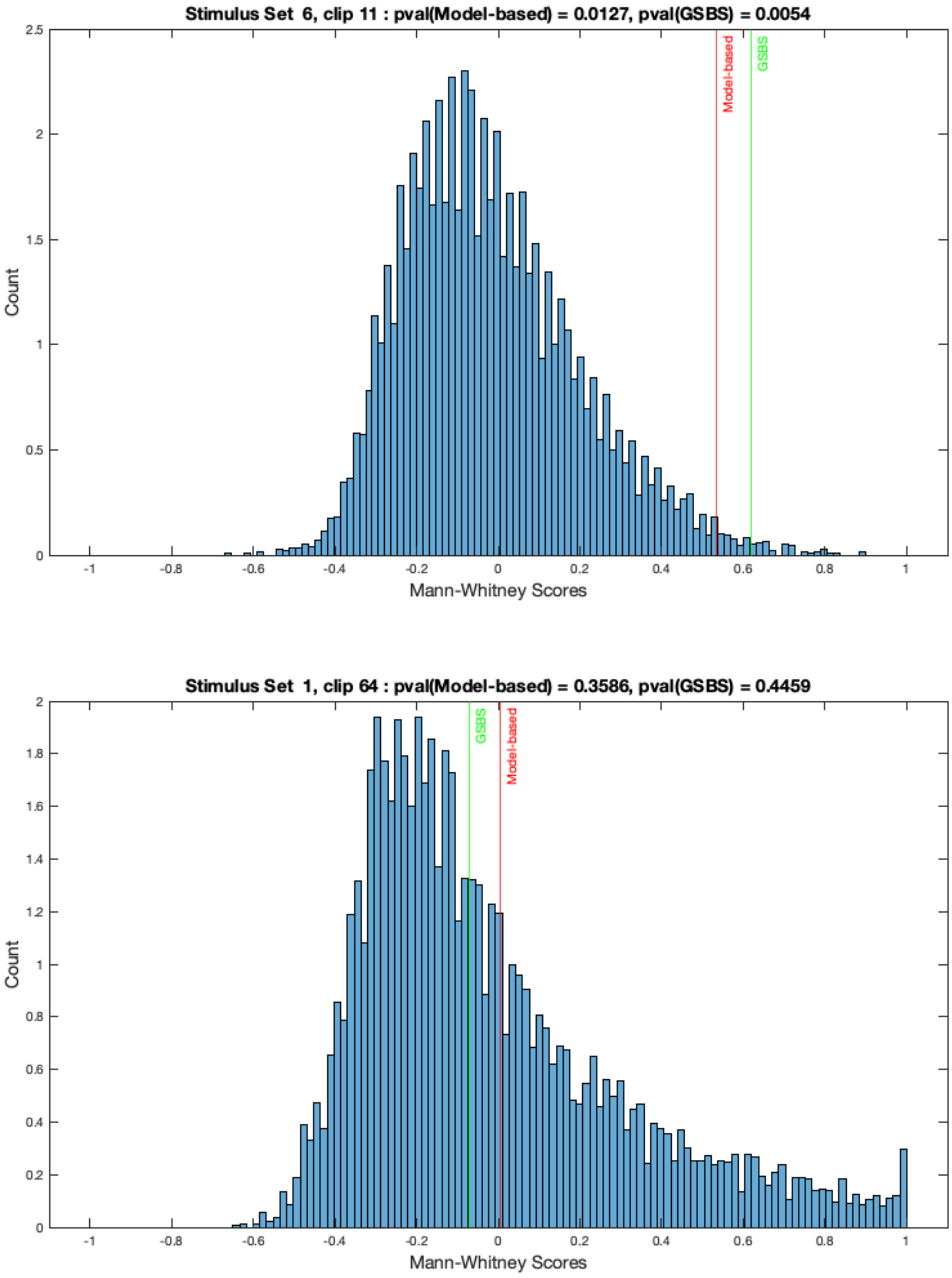


Figure 4. Mann-Whitney scores distribution for two clips. Histograms representing the surrogate null distribution of Mann-Whitney scores between D01 and D02 (*mwstat0*) for two example clips taken from different stimulus sets (01 and 06). Red and green lines represent the Mann-Whitney scores computed between D1 and D2 (*mwstat*) for model-based and GSBS respectively. The upper panel depicts an instance in which both approaches appear to be significant (indicating a difference from the null distribution), whereas the lower panel depicts a pair of non-significant comparisons, where the human-like null would not be rejected.

Results

Event Annotations by Method

To understand whether the annotations of human participants and our two methods (GSBS and model-based) exhibited similar patterns, we looked at each segmentation process separately and extracted its corresponding summary statistics. In addition, we also examined the effect of two of the most prominent features of our videos on the resulting number of events, namely scene type and clip duration. More specifically, for all methods, to assess whether the average number of event boundaries was significantly different across scene type and clip duration, we ran a linear mixed-effect model with scene type (3 levels) and clip duration (6 levels) as fixed effect and participant/stimulus set as random-effect (coded as random intercept). The model was fitted on data solely containing clips longer than 10 seconds and, in case of significant main effects, we conducted a contrast analysis across factor levels via the *estimate_contrasts()* function of the *modelbased* R package (Makowski et al., 2020). All models were fitted with the *lmerTest* R package (Kuznetsova et al., 2017) and significance values were obtained via the Satterthwaite's method (Satterthwaite, 1946).

Human Event Segmentation

We first looked at the event annotations provided by the 131 human participants that completed the online experiment and passed our exclusion criteria for clips longer than 10 seconds in 13 stimulus sets. As a result of our exclusion criteria, the model revealed a significant main effect of both video duration ($F(5,4850.6) = 423.15$, $p < .001$) and scene type ($F(2,4845.3) = 941.66$, $p < .001$) with the number of event boundaries increasing for longer clips (see Table S1, Supplementary Materials) and for scenes containing more perceptual content. More specifically, in our data, human observers in the online experiment provided more annotations in the city-based clips (M = 8.76; SE = 0.33) than in both campus/countryside (M = 5.47; SE = 0.24) and office/café ones (M = 4.47; SE = 0.11). Contrast analysis revealed that all pairwise differences across scene type levels were statistically significant with city clips featuring more event boundaries than both campus/countryside and office/café clips (see Table S5). Mean number of reported event boundaries by scene type are shown in Figure 5C.

GSBS Event Segmentation

We then examined the event boundaries provided by the GSBS segmentation algorithm (Geerligs et al., 2021) that was fitted on the EEG data of 13 participants watching naturalistic movies and providing estimates of duration. Coherently with the model fitted on human annotations from the online experiment, there was a significant main effect of video duration (*F*(5,476.56) = 725.16, *p* < .001), however there was no significant effect of scene type on the number of events extracted via GSBS (*F*(2,475.62) = 0.4553, *p* = 0.634). More specifically, the segmentations generated by GSBS featured roughly the same number of event boundaries for city-based (*M* = 9.72; *SE* = 0.45), campus/countryside (*M* = 9.56; *SE* = 0.40) and office/café clips (*M* = 9.64; *SE* = 0.33). Mean number of GSBS's event boundaries by scene type are shown in Figure 5B.

Model-based Event Segmentation

We then focused on the event boundaries generated by our model-based approach that processed the same 13 sets of naturalistic clips watched by both online raters and participants completing the EEG study. Consistent with the model fitted on human annotations from the online experiment, there was a significant main effect of both video duration (*F*(5, 483.42) = 89.77, *p* < .001) and scene type (*F*(2, 483.87) = 11.09, *p* < .001). More specifically, the number of event boundaries extracted by our model-based approach was greater for campus/countryside (*M* = 8.81; *SE* = 0.40) than for city-based clips (*M* = 8.21; *SE* = 0.39) and office/café ones (*M* = 7.19; *SE* = 0.28). However, as revealed by contrast analysis, only office/café clips featured a statistically significant lower number of boundaries than both city and campus/countryside videos (see Table S6). Mean number of model-based generated event boundaries by scene type are shown in Figure 5A.

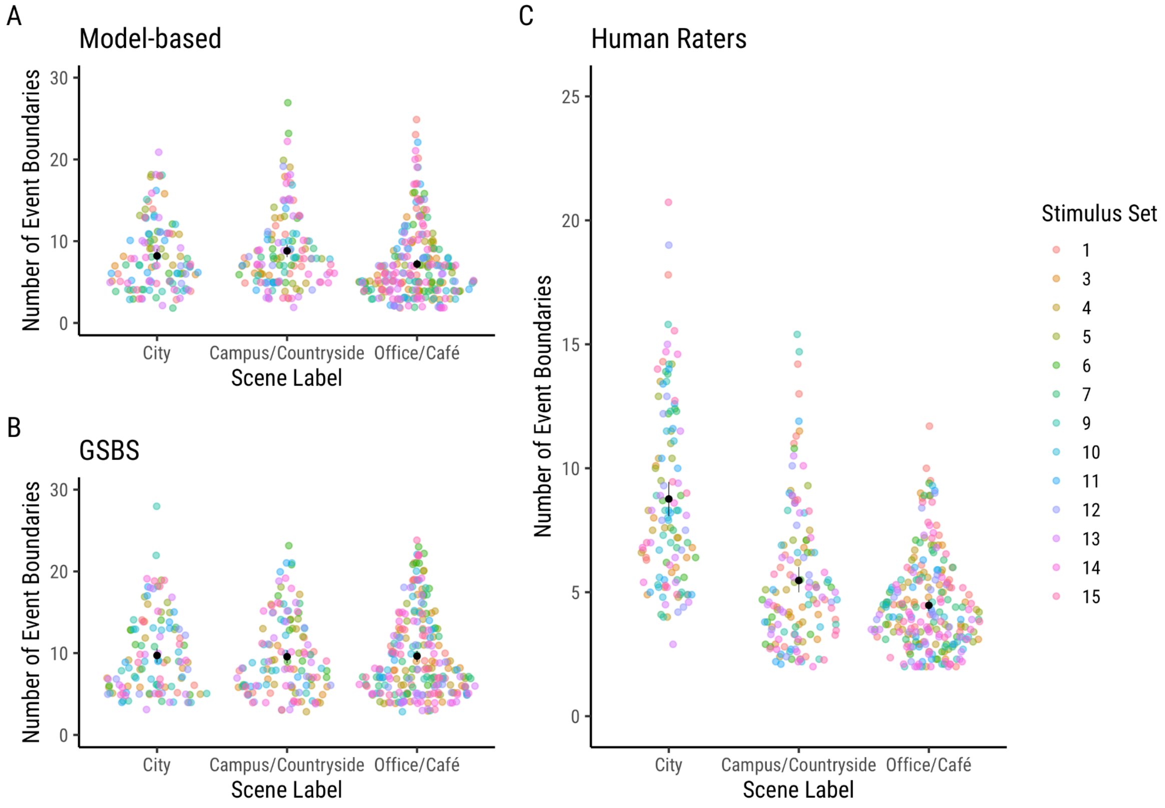


Figure 5. Number of Event Boundaries across Models by Scene Types. Sina plots representing the difference between the estimated number of event boundaries for each approach (model-based, GSBS, human) between different scene types. Each point represents the number of estimated boundaries for a single clip, coloured based on the stimulus set to which it belongs (13 stimulus sets after exclusion of set number 2 and number 8). Error bars represent the bootstrapped 95% confidence interval of the mean. A. Model-based approach (our computational model); B. GSBS fitted on participants' EEG data. C. Human Raters that completed the experiment online.

Comparing Event Segmentations Provided by Different Methods

To understand how the different annotation methods performed - specifically whether the annotations provided by the GSBS data-driven approach or the model-based approach corresponded to what human raters provided during the online experiment - we compared the model-based and GSBS-based annotations with human raters as a ground truth. To do so, as reported in Methods, we created two methods based on different distance measures and statistical approaches: a Direct and a Meta approach.

Direct Approach comparing human, data-driven, and model-based raters

The Direct approach is based on a measure of distance operationalised as the mean squared distance between point-processes (different annotation/segmentation processes). Computing a single value of distance for each possible combination of participant and scene type allowed us to construct a null distribution representing the space of potential distances between human raters. By then calculating the same metric between the single annotations provided by each approach (data-driven and model-based) and the ones generated by human raters and comparing it with the null distribution, we were able to assess whether the average distance between GSBS, model and human raters was significantly greater than between human-raters.

Human raters versus GSBS (EEG) rater

We compared the average of the GSBS-human rater distances with a distribution composed of the inter-rater distances between segmentations carried out by human raters online. The proportion of distances between human raters (p-value) that was more extreme than the mean of GSBS-observers distances was lower than α (0.05) only 4 times out of the 39 possible combinations (10.25% - Table 1) between stimulus set and scene type (13x3). Out of these 4 instances, 3 were related to clips shot in offices/cafés and 1 to videos filmed around the University of Sussex campus and/or countryside (see Table 1). An example of the distribution of inter-rater distances for each scene type for two stimulus sets is provided in Figure 6.

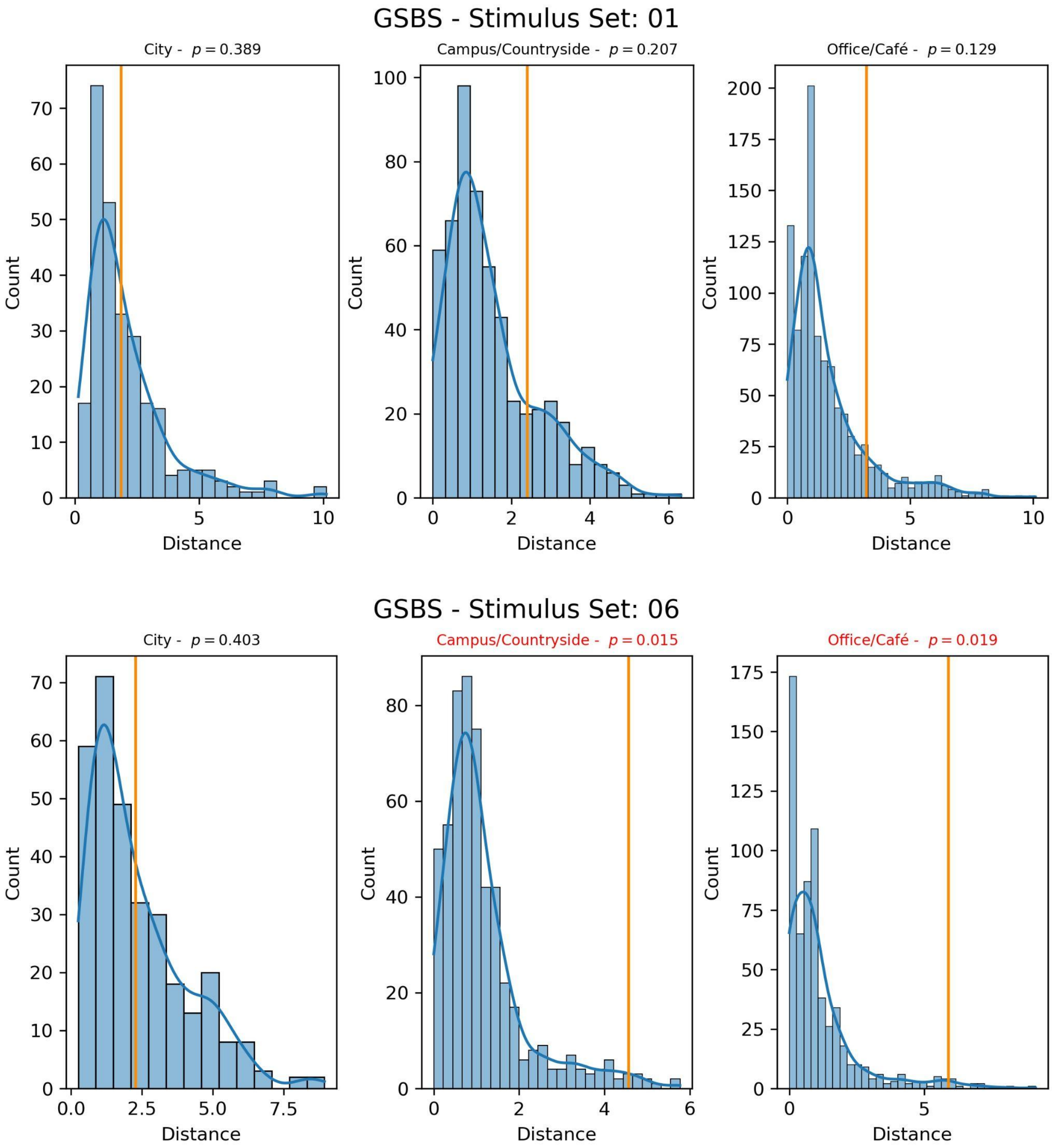


Figure 6. Inter-rater distances between human raters and between humans and GSBS. Histograms representing the distributions of inter-rater distance (in blue) for each scene type (city, campus/countryside, office/café) for two example stimulus sets. Orange lines correspond to the average distance value between the GSBS-produced segmentations and human raters for clips of each specific stimulus set and scene type. Red plot labels highlight examples of significant instances (wherein the average distance from the GSBS-produced event segmentations to the human raters was extreme).

Human raters versus model-based rater

To compare the performance of our model-based approach, applied by scene type, we compared the average of the model-based rater vs human raters distances with a distribution composed of the inter-rater distances between segmentations carried out by human observers online. The proportion of distances between human raters (p-value) that was more extreme than the mean of the model-human distances was lower than α (0.05) 2 times out of the 39 possible combinations (5.12% - Table 1) between stimulus set and scene type (13x3). These 2 instances corresponded to videos filmed around the University of Sussex campus and/or countryside (Table 1). Figure 5 shows the distribution of inter-rater distances by scene type for the same stimulus sets reported for GSBS (Figure 7).

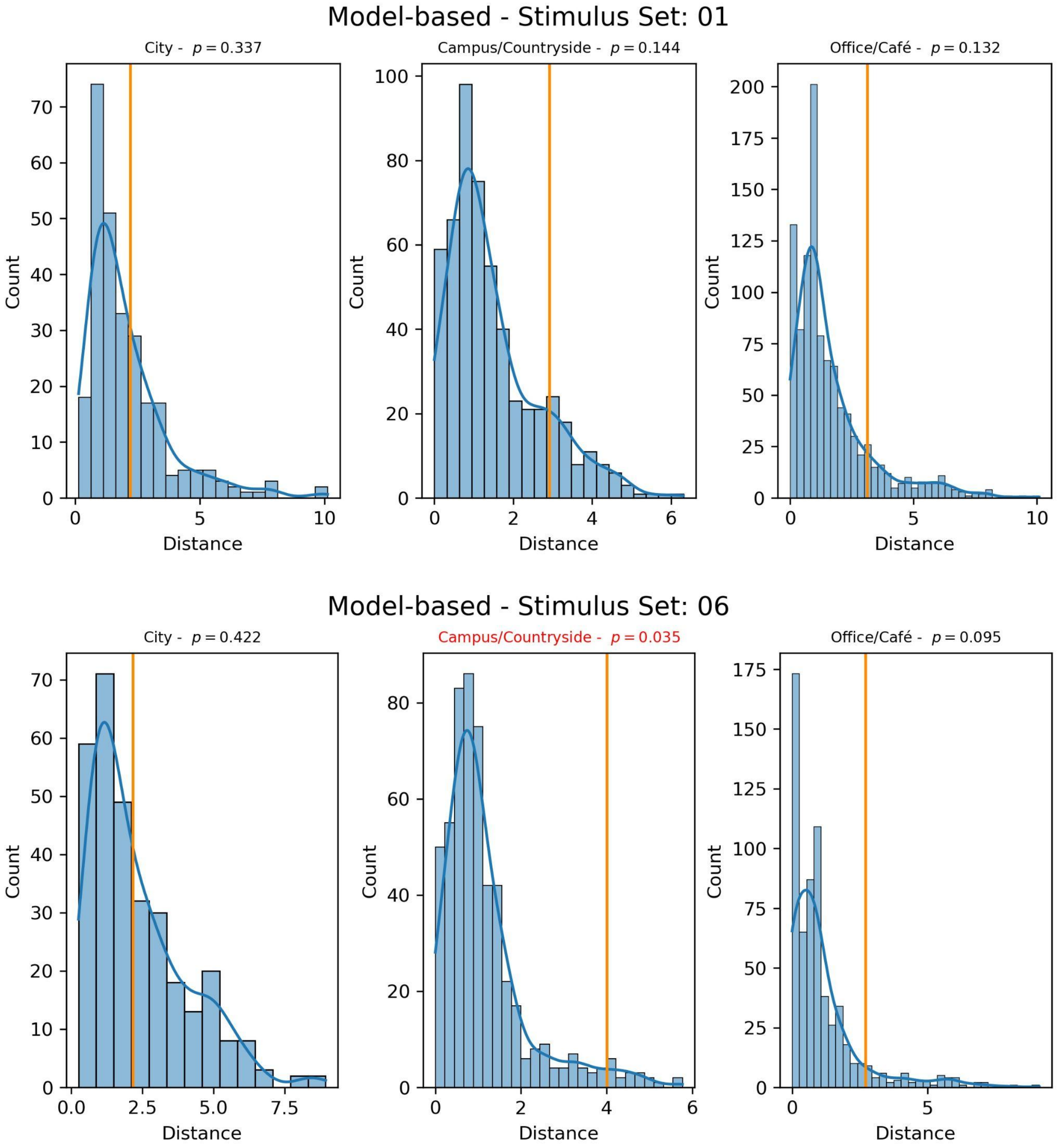


Figure 7. Inter-rater distances between human raters and between humans and model-based approach. Histograms representing the distributions of inter-rater distance values between human raters (in blue) for each

scene type (city, campus/countryside, office/café) for two example stimulus sets. Orange lines correspond to the average distance between the model-based segmentations and the human raters for the clips of each specific stimulus set and scene type. Red plot labels highlight examples of significant instances (wherein the average distance from the model-produced event segmentations to the human raters was extreme).

Control Analysis: Random raters vs Human raters

The control analysis was needed in order to make sure that both GSBS and our model-based approach didn't produce segmentations that were not significantly different from the ones produced by human raters because all raters were simply producing a random series of annotations. To do so, we compared the average of the human-random raters' distances with a distribution composed of the inter-rater distances between segmentations carried out by the original human observers. As expected, the proportion of inter-rater distances that was more extreme than the mean of human-random raters' distances was lower than α (0.05) in 34 instances out of 39 possible combinations (87.17% - Table 1). Albeit only dependent on clip duration rather than content (because the random model was not exposed to any content in any sense) out of the 34 instances, 9 corresponded to clips shot in the city, 12 to the ones recorded around campus and/or countryside and 13 to videos captured in an office or café (Table 1). Therefore we can conclude that if our other approaches - including humans - were simply generating event segmentations at random, we should have been able to detect this by observing a similarly high percentage of significant instances.

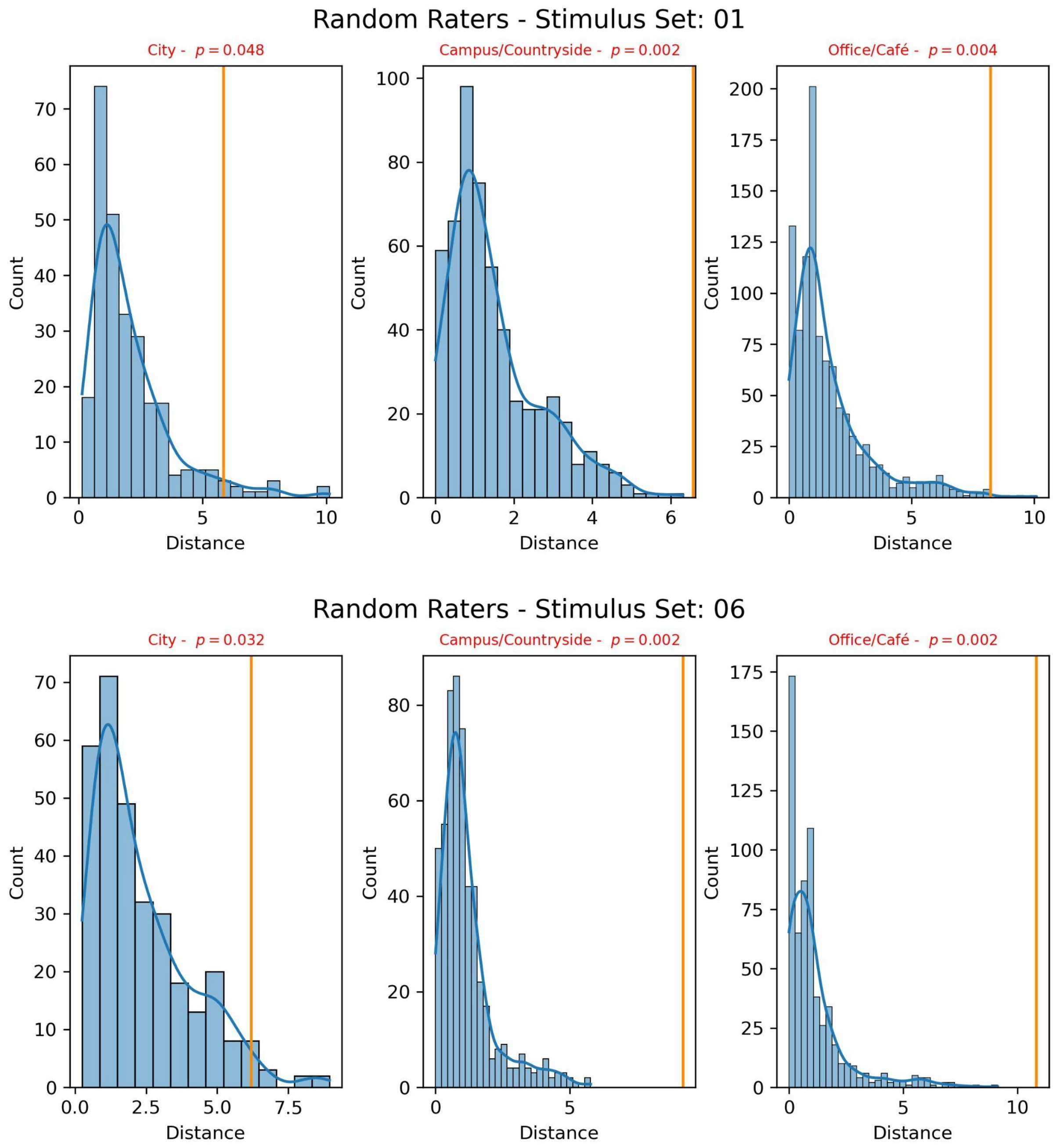
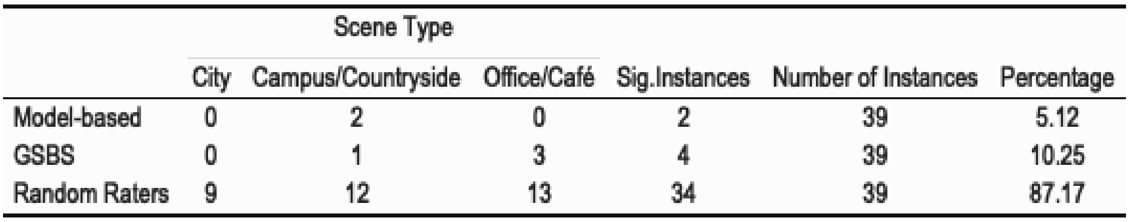


Figure 8. Inter-rater distances between human raters and between humans and random raters. Histograms representing the distributions of inter-rater distance between human raters (in blue) for each scene type (city, campus/countryside, office/café) for two stimulus sets. Orange lines correspond to the average distance value between the random raters' segmentations and the human raters for the clips of each specific stimulus set and scene type. Red plot labels highlight examples of significant instances (wherein the average distance from the random-generated event segmentations to the human raters was extreme).



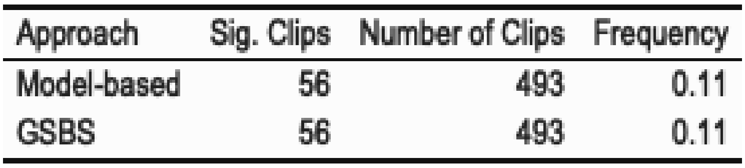
	Scene Type					
	City	Campus/Countryside	Office/Café	Sig.Instances	Number of Instances	Percentage
Model-based	0	2	0	2	39	5.12
GSBS	0	1	3	4	39	10.25
Random Raters	9	12	13	34	39	87.17

Table 1. Direct Approach - Performance Summary over Methods across Scene Types. Table summarising the number and percentage of significant instances (significantly different from human raters) computed by our Direct Approach across different methods/raters and divided by scene type.

Meta Approach comparing human, data-driven, and model-based raters

In our Meta approach, to assess the similarity between the event boundaries reported by humans online and the ones generated by either the GSBS algorithm applied to participants' EEG or by our model-based approach, we proceeded in three steps. First, we computed a Mann Whitney statistic value (*mwstat*) by assessing the stochastic dominance between D1 (the set of 10 average minimum distance values computed between the event boundaries of our model of choice and humans) and D2 (the set of 45 average minimum distance values computed across all human annotations). Second, we generated a surrogate null distribution of Mann Whitney test statistics by repeating the same procedure 10000 times per clip and by substituting the original models' segmentations with Poisson random generated boundary sequences (*mwstat0* - see the "Meta Approach: Average Minimum Distance and Mann-Whitney test" section in Methods for more details). Finally, we computed a single p-value clip by clip by checking whether each *mwstat* belonged to its corresponding null distribution *mwstat0*. More specifically, if the proportion of *mwstat0* values that were more extreme than *mwstat* was very low (e.g. below a threshold of 0.05 divided by number of comparison as per Bonferroni correction), then we could conclude that, for that specific clip, either model-based rater or GSBS rater didn't segment the video like a human rater.

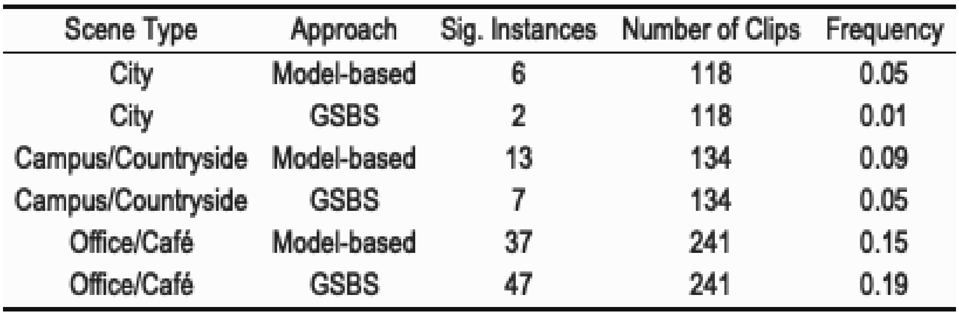
As reported in Table 2, for both model-based and GSBS methods, across all clips, the proportion of Mann Whitney test statistics that didn't belong to their corresponding surrogate null distribution and that remained significant after correction for multiple comparison was 56/493 (11.3%).



Approach	Sig. Clips	Number of Clips	Frequency
Model-based	56	493	0.11
GSBS	56	493	0.11

Table 2. Number of Significant Clips (corrected for multiple comparison). Table including the number and proportion of significant clips (significantly different from human raters) computed by our Meta Approach for both our model-based and GSBS approach.

When dividing clips by scene type, GSBS marginally outperformed the model-based approach for both city (GSBS: 2/118; model-based: 6/118) and countryside videos (GSBS: 7/134; model-based: 13/134), while the converse was true for office scenes (GSBS: 47/241 ; model-based: 37/241 - Table 3).



Scene Type	Approach	Sig. Instances	Number of Clips	Frequency
City	Model-based	6	118	0.05
City	GSBS	2	118	0.01
Campus/Countryside	Model-based	13	134	0.09
Campus/Countryside	GSBS	7	134	0.05
Office/Café	Model-based	37	241	0.15
Office/Café	GSBS	47	241	0.19

Table 3. Number of Significant Clips across Scene Types (corrected for multiple comparison). Table including the number and proportion of significant clips (significantly different from human raters) computed by our Meta Approach for both our model-based and GSBS approach across scene types.

The same qualitative patterns albeit with approximately a 3-fold increase in number of significant instances emerged when considering the uncorrected results (see Table S6 and S7 in Supplementary Material).

Discussion

This study examined how humans and different computational approaches segment continuous, like-life, naturalistic videos into events. Specifically, we assessed whether event segmentation processes generated by a process based on EEG data (GSBS), and a model of event segmentation developed to model human time perception, could be reasonably said to come from a sample of human raters who segmented the same videos in an online task. Our main finding was that both GSBS and our model-based approach were able to generate event boundaries that were similar to ones reported by our independent sample of raters. This was confirmed by two different inferential approaches: the first relying on inferences made on aggregated mean-squared distance values between segmentations and the second based on the average minimum distance across vectors of event boundaries.

The “Direct” approach demonstrated that, across 39 possible instances (13 participants by 3 scene types), both the GSBS and model-based approach exhibited a low number of cases wherein the average mean-squared distance between their generated segmentations and human provided annotations was an outlier in the distribution of human-human distances (Table 1). Conversely, when the same approach was applied considering a series of 100 randomly generated raters, the number of significant instances rose to 34/39, indicating that our Direct method could clearly discriminate informed segmentation processes from structured, but randomly generated segmentation sequences.

Our second, “Meta” approach resulted in similar results, with GSBS and model-based approach performing equally well (11.3% of significant instances overall), but with different qualitative patterns across scene types: GSBS was better with city and countryside videos, but worse for office ones (see Table 2 and 3).

Overall, using two different measures of difference and two different inferential processes, our results support the idea that both GSBS based on EEG (Geerligs, et al., 2021) and our model-based approach (Fountas et al., 2022; Roseboom et al., 2019) do a reasonably good job of approximating human–like event segmentations for naturalistic video - neither approach stands out as any more different from a sample of human raters providing event boundaries than any individual human rater.

Event segmentation models

Despite the central role of event segmentation for human memory and generalisation and the vast number of papers released in recent years (e.g., see Bird, 2020; Shin & DuBrow, 2021 for review), few studies have attempted to build explicit computational models of event segmentation. One of the first modelling attempts was made by Reynolds, Zacks and Braver (2007) who attempted to formalise the guiding principles of Event Segmentation Theory (EST; Zacks et al., 2007) by means of a gated recurrent neural network inspired model (Hochreiter & Schmidhuber, 1997). The model leveraged prediction errors between consecutive frames of simple 3D motion captures videos as the central gating mechanism to generate event boundaries, update current event models, and predict the next time step of the input. More recently, Elman & McRae (2019) built a computational model of event knowledge by means of a simple recurrent neural network coupled with localist representations (e.g., one concept = one unit) to model the generation of conceptual structures and schemas and producing correct inferences based on linguistic stimuli.

Another recent attempt to formally model event segmentation mechanisms comes from Franklin and colleagues (2020) who created the Structured Event Memory Model (SEM). This model aimed at replicating the full spectrum of human event segmentation mechanisms and related functions (learning, inference, prediction and memory). The model overcomes the limitations of previous formalisations by defining three hierarchical main components. First, individual scenes (e.g., the actual context/description of the environment composed of objects and the relations between them) were defined and modelled as (distributed) holographic reduced representations (Plate, 1995) with a variational autoencoder used to learn and provide the representational space (Kingma & Welling, 2014) to the model. Scenes were then clustered into events by means of a sticky Chinese Restaurant Process (sticky-CRP; Fox et al., 2011), a non-parametric Bayesian clustering algorithm that made sure that timepoints were assigned to specific events based on the frequency and recency of previous events and that acted, in parallel, as constraint for the definition of scenes themselves (as would typically happen for schema scripts used by humans). Finally, the dynamics within individual events were instantiated by means of a 4-layer neural network with gate recurrent units (GRU; Cho et al., 2014) as reported in the aforementioned models. Crucially, SEM was able to match human segmentations in a variety of tasks ranging from to 3D motion captures to fully naturalistic videos (i.e., a person washing dishes).

The main differences between the above-described models and our model-based approach are twofold. First, despite their differences, all of the above models rely - at some level of their architecture - on gated recurrent neural networks to model event dynamics and to account for memory within the system. Second, all of them were built to *explicitly* model human-like mechanisms of event segmentation. Our computational model i) didn't feature any form of recurrency and ii) was originally created to reproduce human estimates of duration (Roseboom et al., 2019) by simply leveraging a feedforward deep convolutional neural network (previously AlexNet, currently MobileNetV2 - Krizhevsky et al., 2017; Sandler et al., 2019) that was trained on object classification. When considering both its training regime and the absence of any additional components or mechanisms mimicking memory and/or the capacity to generate high-order level relations between events, the results that we obtained in this current study appear to be even more encouraging. The fundamental assumption and thread that links all the studies built around this modelling approach is the intuition presented in the initial paper of this series by Roseboom and colleagues (2019): that the experience of time might emerge as by-product of the accumulation of salient events (event boundaries) in biological and/or artificial perceptual classification networks. In this context, event boundaries are generated in correspondence to large prediction errors (relatively large deviations from the prior state of the network).

Our rationale is supported by recent work from Kumar et al. (2022) who found that the generation of event boundaries during story listening was associated with increases in transient Bayesian surprise, operationalised as normalised KL divergence between the probability distributions of successive words computed by the deep learning language model GPT-2, but not with simpler prediction error measures. This is also coherent with recent developments in our model architecture that, following hierarchical predictive coding principles, was extended by the inclusion of a generative model component aimed at mimicking episodic and semantic memory and that was able to correctly estimate durations for retrospective time judgments (Fountas et al., 2022).

All the models discussed in this section, including our model, required the definition of a series of specific parameters (e.g., in our case related to the model attentional threshold). However, the GSBS model fitted on participants' EEG data is a fully data-driven model that only requires the definition of the maximum number of event boundaries that can be retrieved in the timeseries of interest (Geerligs et al., 2021). Similar data-driven approaches have typically been used to extract latent states from high-dimensional brain structures (e.g. fMRI signals related to specific regions or networks) of participants during naturalistic stimulation (e.g. movies, audio-narratives, music pieces - Baldassano et al., 2017; Cohen et al., 2022; Williams et al., 2022).

All of these approaches leverage the fundamental assumption that event boundaries detected by participants seem to co-occur with shifts in stable patterns of neural activity and with increased hippocampal activity (Baldassano et al., 2017; Ben-Yakov & Henson, 2018). Among all, the most popular methods to accomplish this task are probably those based on Hidden Markov Models (HMM). HMM has been shown to be able to retrieve neural states at different hierarchies corresponding with the events reported by participants when segmenting different kinds of content. Examples range from adults (Baldassano et al., 2017) or children (Cohen et al., 2022) watching movies while undergoing fMRI or EEG (Silva et al., 2019) to the detection of events across different brain regions matching participants' phenomenological segmentation of different musical excerpts (Williams et al., 2022). Both

HMM and GSBS are not designed to identify recurrent states (the first time point is always within event 1 and the last one with event k), however they differ in terms of both the selected measure of distance and fitting time. Crucially, HMM's optimal number of states is defined as the number of states with the largest average difference of within- versus across-state correlations, whereas GSBS' t-distance uses the t-statistic of the difference between states that are consecutive.

GSBS has shown to outperform HMM in retrieving sequences of events from both simulated and real fMRI data (Geerligs et al., 2021). However, while GSBS has typically been fitted on group-averaged ROI-based fMRI data (Geerligs et al., 2022), our current results demonstrate that the model was able to retrieve meaningful series of event boundaries also when fitted on single-subject EEG data. More specifically, it's particularly encouraging that its performance was very close to the one of our model when considering that i) their input was radically different (video frames for our model, electroencephalographic data for GSBS) and also ii) that GSBS input data corresponded to normalised full scalp sensor-space data without source reconstruction and/or spatial filtering (aside from ICA-based cleaning during preprocessing).

Clarifications and Future Directions

Building on the results provided in Fountas et al. (2022), the results presented here provide a clear demonstration of the power of our model-based approach to event segmentation, developed initially in the domain of time perception. However, the data used to arrive at these inferences could be made more ideal.

First, both of our presented analysis approaches (Direct and Meta) target a different (albeit similar) question to that considered in previous work in the event segmentation literature. Specifically, we consider whether the event boundaries generated by a model (or brain data-driven approach) are "human-like" enough to be considered to fit within a distribution of human-provided segmentation sequences. Other studies instead addressed whether the segmentation sequences obtained from a specific model (e.g., HMM) matched the annotations provided by human raters by means of different two-steps procedures. First, a measure of similarity/distance is typically defined: either number of matching boundaries between sequences within a pre-specified time-window (e.g., 3 seconds - Baldassano et al., 2017; Williams et al., 2022), the point-biserial correlation between different segmentations (Franklin et al., 2020; Zacks et al., 2006) or the Jaccard index to estimate the similarity between two time-series (Geerligs et al., 2021), etc (e.g., see Cohen et al., 2022; Lee et al., 2021). Second, the selected metric is typically re-computed many times within a permutation-based approach (e.g., 1000) by shuffling the event boundaries of one of the two sequences in time and by keeping both the number of boundaries and the distance between them constant. The original measure of distance/similarity is then compared to this permuted random distribution to obtain a statistical inference. As illustrated by these example methods, there are many possible measures of similarity between two sequences of event boundaries (whether human or model generated), and potential methods with which to generate an inference to conclude the measures indicate similarity. Decisions about distance measures and inferential processes contain different assumptions and target different versions of the desired question. Other researchers may prioritise, and previously have prioritised, different aspects or questions than we have.

Second, our inferences were based on a limited number of participants undergoing the EEG study (n=13) watching several, diverse clips without repetitions (80 per participant). Both the low number of participants and the absence of repetitions of the same set of stimuli across participants could be considered as a drawback, especially for GSBS. As reported previously, the signal-to-noise ratio of the input given to GSBS could be increased by feeding the model specific components or sources rather than sensor-space data. In this context, the collection of at least 10 different human segmenters per stimulus set, together with the development of two novel statistical approaches and distance measures which are robust to small sample sizes, represents two important ways to mitigate these problems. Future studies could attempt to reduce the impact of these issues by re-running the same EEG experiment with a larger number of participants watching a standard set of clips. This will allow the application of inter-subject correlation analysis (ISC - Hasson, 2004; Hasson et al., 2008) by leveraging statistical approaches such as Canonical Correlation Analysis (CCA - Hotelling, 1936) or Correlated Component Analysis (CorrCA - Parra et al., 2019) which are able to linearly combine the EEG signal coming from different participants to extract the components that are maximally correlated among them and that typically correspond to the neural processing of different features of the stimuli or different cognitive processes (e.g., Cohen & Parra, 2016; Dmochowski et al., 2012). This will in turn allow the possibility to input GSBS with EEG components actually reflecting common processing of the naturalistic videos and that, upon reprojection or additional source reconstruction (e.g. see Haufe et al., 2014) will also be biologically interpretable. New questions would then be able to emerge: in the case of multimodal videos (e.g., video plus sound), do components that are maximally correlated across participants but reflect different modalities generate different event segmentations?

Of note, in comparison to related research on the topic, our study used relatively short (1-64 s) silent video clips instead of long and complex audiovisual stimuli like movies or TV shows thereby being limited i) to not fully ecological stimuli in this aspect and ii) to events that pertain to relatively short timescales. While we recognise that the absence of sound could constitute a problem for generalising inferences about the obtained segmentations, the idea of using naturalistic clips of life-like situations instead of complex edited materials constitutes, as claimed in the Introduction, one of the theoretical pillars of our study. The use of explicitly edited audiovisual content, in which each transition and change of perspective is introduced for specific communicative reasons (e.g. see Grall & Finn, 2022) risks to turn the question of "How do people segment their phenomenological experience into events?" into "How do people *detect* events and boundaries?".

In our view, instead of adopting movies and/or excerpts from TV, a more ecological and productive approach would be to use non-edited clips filmed in first-person perspective in naturalistic contexts which truly reflect the snapshots characterising our everyday experience. For instance, in order to study event compression in memory, Jeunehomme et al. (2018) used a series of clips recorded in first-person perspective by participants themselves while they completed a series of typical daily actions on campus. Clips were similar to the ones filmed on campus that we used in the current task with the sole differences that they were presented frame-by-frame and that they featured way more first-person perspective actions and interactions than ours. We think that rather than approaching the question from either a "top-down" (e.g. using complex edited materials) or a "bottom-up" perspective (e.g. using sequences of images and/or words separated by artificial boundaries), using multimodal (i.e. audiovisual) naturalistic clips shot in first-person perspective that represent daily life scenarios (including actions and social interactions) constitutes a more proficient and ecologically valid way forward in the study of how people segment experience.

Finally, it might be argued that the deep convolutional neural network that constituted the centrepiece of our computational model is not sufficiently biologically plausible, or in some way lacks appropriate training data for the employed purpose. There is certainly an active discussion in neuroscience and computer vision regarding the best way to characterise primate vision and whether feed-forward convolutional networks are useful for this and related purposes (e.g., Kriegeskorte, 2015; Lindsay, 2021 but see Bowers et al., 2022). While these discussions are essential, they are to some degree beside the point for our present work. As discussed in Roseboom et al. (2019), Sherman et al. (2022) and Fountas et al. (2022) in relation to modelling time perception, the critical point isn't necessarily whether this specific network is the correct or most biologically plausible one, but simply whether it contains sufficient features to accomplish the task of approximating human event segmentation. Our results show that it does. In addition, the structure of the neural network itself becomes less important in light of the probabilistic interpretation of our model (e.g., the output of the model can be viewed as an approximate posterior distribution over latent states), since representations in the latent space actually map onto pre-defined objects for which humans typically have words and that are very likely to emerge at some stage of the perceptual hierarchy (see Fountas et al., 2022, for further discussion). However, future work might benefit from investigating different model architectures (e.g. recurrent neural networks - Kietzmann et al., 2019; Spoerer et al., 2017) and different training regimes/sets such as the *ecoset* database (https://codeocean.com/capsule/9570390/tree/v1; Mehrer et al., 2021) or the THINGS database (https://things-initiative.org/; Hebart et al., 2019) in an attempt to match for broadly more biologically and ecologically plausible constraints. Whether these different models would perform better would be a matter of empirical interest.

One interesting way to further assess the role of biological plausibility and network architecture on the model segmentation of naturalistic videos would be to compare the performance of models having a different Brain Score (http://www.brain-score.org/), a measure aimed at quantifying the degree to which deep neural networks appear to be functionally similar to the human brain. Brain Score is computed by comparing each network to a series of neural and behavioural benchmarks (Schrimpf, Kubilius, Hong, et al., 2020; Schrimpf, Kubilius, Lee, et al., 2020). Intuitively, a model with a higher Brain Score should generate event boundaries that are more human-like than boundaries extracted from less biologically plausible networks. This idea, together with the general intuition behind our current study, resonate with what Ma & Peters (2020) have argued - that cognitive science would greatly benefit from the use of DNNs as part of models of human behaviour. This perspective was recently endorsed also in the domain of brain sciences via the *neuroconnectionist* research programme (Doerig et al., 2022). More generally, as Guest and Martin (2021a) recently argued, using computational modelling and model-driven hypotheses is a useful, perhaps essential tool for theory building in psychology and cognitive science (see also Guest & Martin, 2021b about potential logical and/or inferential fallacies that might arise when comparing the human behaviour and/or brain activity with DNNs).

Conclusions

In this study, we compared event annotations provided by two different methods with those provided by human participants. The two approaches come from different research directions, with the data-driven GSBS method being developed in the context of event segmentation literature, and the model-based approach developed in the context of modelling human time perception. When compared to human-provided event annotations in naturalistic videos, we found that both do a good job of broadly matching the properties of human raters - specifically, neither method stood out relative to a sample of human raters. This work goes some way towards reconciling the fields of time perception and episodic memory by demonstrating that we are apparently studying the same fundamental units of processing in experience - “events” as segmented from continuous experience. Further studies are required to fully characterise the nature of “events” and of the event segmentation process, but this work provides initial evidence for a potential full consolidation of these research fields.

References

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2021). Behavioral, Physiological, and Neural Signatures of Surprise during Naturalistic Sports Viewing. *Neuron*, *109*(2), 377-390.e7. https://doi.org/10.1016/j.neuron.2020.10.029

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. A., & Kosie, J. E. (2021). How Does the Mind Render Streaming Experience as Events? *Topics in Cognitive Science*, *13*(1), 79–105. https://doi.org/10.1111/tops.12502

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

Bird, C. M. (2020). How do we remember events? *Current Opinion in Behavioral Sciences*, *32*, 120–125. https://doi.org/10.1016/j.cobeha.2020.01.020

Blinn, K. A. (1955). Focal anterior temporal spikes from external rectus muscle. *Electroencephalography and Clinical Neurophysiology*, *7*(2), 299–302. https://doi.org/10.1016/0013-4694(55)90043-2

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review*, *4*(2), 184–197. https://doi.org/10.3758/BF03209393

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F. G., Hummel, J., Heaton, R. F., Evans, B., Mitchell, J., & Blything, R. (2022). *Deep Problems with Neural Network Models of Human Vision*. PsyArXiv. https://doi.org/10.31234/osf.io/5zf4s

Boylan, C., & Doig, H. R. (1989). Effect of saccade size on presaccadic spike potential amplitude. *Investigative Ophthalmology and Visual Science*, *30*(12), 2521–2527. Scopus.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bromis, K., Raykov, P. P., Wickens, L., Roseboom, W., & Bird, C. M. (2022). The Neural Representation of Events Is Dominated by Elements that Are Most Reliably Present. *Journal of Cognitive Neuroscience*, *34*(3), 517–531. https://doi.org/10.1162/jocn_a_01802

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Chollet, F. & others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Clewett, D., Gasser, C., & Davachi, L. (2020). Pupil-linked arousal signals track the temporal organization of events in memory. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17851-9

Cohen, S. S., & Parra, L. C. (2016). Memorable Audiovisual Narratives Synchronize Sensory and Supramodal Neural Responses. *ENeuro*, *3*(6). https://doi.org/10.1523/ENEURO.0203-16.2016

Cohen, S. S., Tottenham, N., & Baldassano, C. (2022). Developmental changes in story-evoked responses in the neocortex and hippocampus. *ELife*, *11*, e69430. https://doi.org/10.7554/eLife.69430

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods,* *Instruments, & Computers*, *34*(4), 613–617. https://doi.org/10.3758/BF03195489

Cutting, J. E. (2014). Event segmentation and seven types of narrative discontinuity in popular movies. *Acta Psychologica*, *149*, 69–77. https://doi.org/10.1016/j.actpsy.2014.03.003

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

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, *207*, 116117. https://doi.org/10.1016/j.neuroimage.2019.116117

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, *140*(4), 552–572. https://doi.org/10.1037/a0023885

Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated Components of Ongoing EEG Point to Emotionally Laden Attention – A Possible Marker of Engagement? *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00112

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., Kording, K., Konkle, T., Van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2022). *The neuroconnectionist research programme* (arXiv:2209.03718). arXiv. https://doi.org/10.48550/arXiv.2209.03718

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology. General*, *142*(4), 1277–1286. https://doi.org/10.1037/a0034024

DuBrow, S., & Davachi, L. (2014). Temporal memory is shaped by encoding stability and intervening item reactivation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(42), 13998–14005. https://doi.org/10.1523/JNEUROSCI.2535-14.2014

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252–291. https://doi.org/10.1037/rev0000133

Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., & Roseboom, W. (2022). A Predictive Processing Model of Episodic Memory and Time Perception. *Neural Computation*, *34*(7), 1501–1544. https://doi.org/10.1162/neco_a_01514

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, *5*(2A), 1020–1056. https://doi.org/10.1214/10-AOAS395

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. *Psychological Review*, *127*(3), 327–361. https://doi.org/10.1037/rev0000177

Geerligs, L., Gerven, M. van, Campbell, K., & Güçlü, U. (2021). A nested cortical hierarchy of neural states underlies event segmentation in the human brain. *BioRxiv*, 2021.02.05.429165. https://doi.org/10.1101/2021.02.05.429165

Geerligs, L., Gerven, M. van, & Güçlü, U. (2020). Detecting neural state transitions underlying event segmentation. *BioRxiv*, 2020.04.30.069989. https://doi.org/10.1101/2020.04.30.069989

Geerligs, L., Gözükara, D., Oetringer, D., Campbell, K. L., van Gerven, M., & Güçlü, U. (2022). A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *ELife*, *11*, e77430. https://doi.org/10.7554/eLife.77430

Geerligs, L., van Gerven, M., & Güçlü, U. (2021). Detecting neural state transitions underlying event segmentation. *NeuroImage*, *236*, 118085. https://doi.org/10.1016/j.neuroimage.2021.118085

Grall, C., & Finn, E. S. (2022). Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience*, *17*(6), 598–608. https://doi.org/10.1093/scan/nsac019

Guest, O., & Martin, A. E. (2021). *On logical inference over brains, behaviour, and artificial neural networks*. PsyArXiv. https://doi.org/10.31234/osf.io/tbmcg

Hasson, U. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced Intersubject Correlations during Movie Viewing Correlate with Successful Episodic Encoding. *Neuron*, *57*, 452–462. https://doi.org/10.1016/j.neuron.2007.12.009

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Wicklin, C. V., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Heusser, A. C., Ezzyat, Y., Shiff, I., & Davachi, L. (2018). Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(7), 1075–1090. https://doi.org/10.1037/xlm0000503

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, *76*(2), 423–434. https://doi.org/10.1016/j.neuron.2012.08.011

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, *28*(3/4), 321–377. https://doi.org/10.2307/2333955

Jeunehomme, O., Folville, A., Stawarczyk, D., Van der Linden, M., & D’Argembeau, A. (2018). Temporal compression in episodic memory for real-life events. *Memory*, *26*(6), 759–770. https://doi.org/10.1080/09658211.2017.1406120

Keren, A. S., Yuval-Greenberg, S., & Deouell, L. Y. (2010). Saccadic spike potentials in gamma-band EEG: Characterization, detection and suppression. *NeuroImage*, *49*(3), 2248–2263. https://doi.org/10.1016/j.neuroimage.2009.10.057

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLOS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 201905544. https://doi.org/10.1073/pnas.1905544116

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What’s new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* [Manual]. https://www.tidymodels.org

Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2022).

Bayesian Surprise Predicts Human Event Segmentation in Story Listening. PsyArXiv. https://doi.org/10.31234/osf.io/qd2ra

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

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *ELife*, *10*, e64972. https://doi.org/10.7554/eLife.64972

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in recording EEGs and event-related potentials II: Source dipoles and source components. *Brain Topography*, *6*(1), 65–78. Scopus. https://doi.org/10.1007/BF01234128

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

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Estimation of model-based predictions, contrasts and means. *CRAN*. https://github.com/easystats/modelbased

Marmor, M. F., & Zrenner, E. (1993). Standard for clinical electro-oculography. *Documenta*

Ophthalmologica, *85*(2), 115–124. Scopus. https://doi.org/10.1007/BF01371127

Matsuo, F., Peters, J. F., & Reilly, E. L. (1975). Electrical phenomena associated with movements of the eyelid. *Electroencephalography and Clinical Neurophysiology*, *38*(5), 507–511. https://doi.org/10.1016/0013-4694(75)90191-1

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8). https://doi.org/10.1073/pnas.2011417118

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

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Parra, L. C., Haufe, S., & Dmochowski, J. P. (2019). Correlated Components Analysis—Extracting Reliable Dimensions in Multivariate Data. *ArXiv:1801.08881 [Cs, Stat]*. http://arxiv.org/abs/1801.08881

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

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

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641. https://doi.org/10.1109/72.377968

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, *6*, 278. https://doi.org/10.3389/fnhum.2012.00278

Pu, Y., Kong, X.-Z., Ranganath, C., & Melloni, L. (2022). Event boundaries shape temporal organization of memory by resetting temporal context. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-28216-9

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. (2014). *Event cognition* (pp. ix, 272). Oxford University Press.

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

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

Roseboom, W., Fountas, Z., Nikiforou, K., Bhowmik, D., Shanahan, M., & Seth, A. K. (2019). Activity in perceptual classification networks as a basis for human subjective time perception. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-018-08194-7

Roseboom, W., Seth, A., Sherman, M., & Fountas, Z. (2022). *The Perception of Time in Humans, Brains, and Machines*. PsyArXiv. https://doi.org/10.31234/osf.io/c7vzx

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *ArXiv:1801.04381 [Cs]*. http://arxiv.org/abs/1801.04381

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

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110. https://doi.org/10.2307/3002019

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

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* (p. 407007). bioRxiv. https://doi.org/10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Sherman, M. T., Fountas, Z., Seth, A. K., & Roseboom, W. (2022). Trial-by-trial predictions of subjective time from human brain activity. *PLOS Computational Biology*, *18*(7), e1010223. https://doi.org/10.1371/journal.pcbi.1010223

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

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences* (pp. xvii, 312). McGraw-Hill.

Silva, M., Baldassano, C., & Fuentemilla, L. (2019). Rapid Memory Reactivation at Movie Event Boundaries Promotes Episodic Encoding. *Journal of Neuroscience*, *39*(43), 8538–8548. https://doi.org/10.1523/JNEUROSCI.0360-19.2019

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

Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science*, *18*(5), 449–455. https://doi.org/10.1111/j.1467-9280.2007.01920.x

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01551

Stephens, G. J., Honey, C. J., & Hasson, U. (2013). A place for time: The spatiotemporal structure of neural dynamics during natural audition. *Journal of Neurophysiology*, *110*(9), 2019–2026. https://doi.org/10.1152/jn.00268.2013

Suárez-Pinilla, M., Nikiforou, K., Fountas, Z., Seth, A. K., & Roseboom, W. (2019). Perceptual Content, Not Physiological Signals, Determines Perceived Duration When Viewing Dynamic, Natural Scenes. *Collabra: Psychology*, *5*(1), 55. https://doi.org/10.1525/collabra.234

Williams, J. A., Margulis, E. H., Nastase, S. A., Chen, J., Hasson, U., Norman, K. A., & Baldassano, C. (2022). High-Order Areas and Auditory Cortex Both Represent the High-Level Event Structure of Music. *Journal of Cognitive Neuroscience*, *34*(4), 699–714. https://doi.org/10.1162/jocn_a_01815

Winkler, I., Debener, S., Muller, K.-R., & Tangermann, M. (2015). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4101–4105. https://doi.org/10.1109/EMBC.2015.7319296

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/10.1038/88486

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., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3–21. https://doi.org/10.1037/0033-2909.127.1.3

Zakharov, A., Guo, Q., & Fountas, Z. (2022). *Variational Predictive Routing with Nested Subjective Timescales* (arXiv:2110.11236). arXiv. https://doi.org/10.48550/arXiv.2110.11236

Supplementary Materials

Computational Modelling: effect of threshold parameters

As reported in the "Computational Modelling" section in Methods, to make sure that the performance of our computational model was not severely influenced by the minimum (T_{min}) and maximum (T_{max}) values of its thresholding mechanism, we re-fitted the model for every possible combination of these parameters in a range between 1 and 10 in steps of 1. The main criterion that was then used to assess the best performing model was to compute the correlation between each clip duration reported by human participants that took part in the EEG study and the duration in seconds estimated by our model-based approach.

In order to generate model estimates of duration, coherently with the regression-based approach of Roseboom et al., 2019, we first fitted a linear mixed effect model per combination of parameters that aimed at predicting the physical duration of each clip solely based on the number of event boundaries generated by our computational model. In this context, the number of events was coded as a fixed effect and the stimulus set was included as a random intercept. The model was subsequently inverted and fed with the number of events to produce a single estimated duration in seconds per clip. Model-based estimated times were then correlated with the ones reported by participants to generate a single correlation value per pair of threshold parameters (Figure S1)

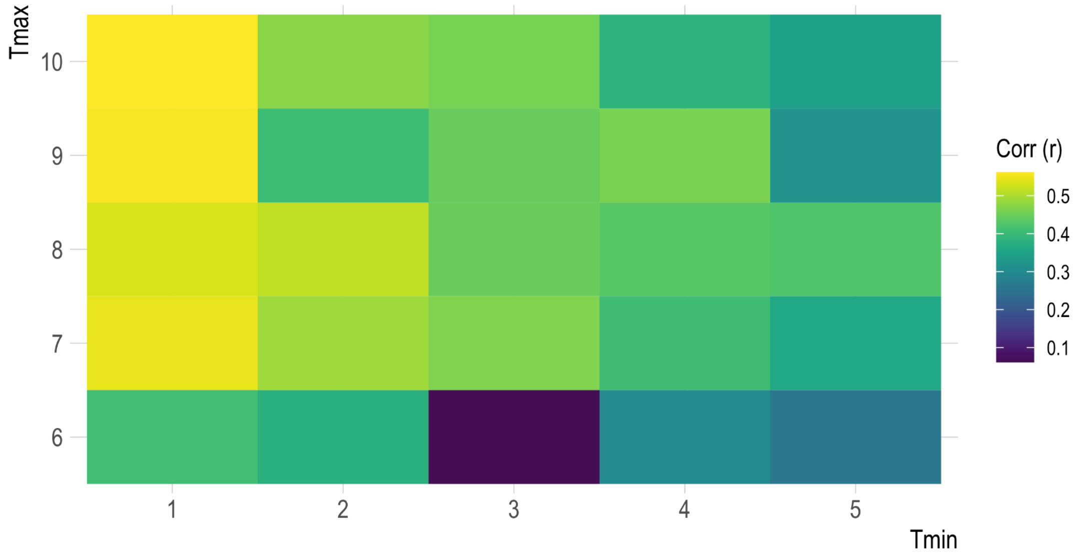


Figure S1. Correlation between human duration estimates and model-based estimates. Heatmap representing the Pearson correlation values computed between the durations provided in seconds by participants taking part in the EEG experiment and the durations estimated by our computational model via regression. Each cell corresponds to a different combination of threshold parameters (Tmin and Tmax).

In addition, we also checked how each instantiation of the model performed in terms of matching the event boundaries included by our sample of human online raters.

Figure S2 represents the number of occurrences across scene types in which the segmentation generated by each model didn't match the corresponding one created by humans. This corresponds to the number of significant instances detected by our "Direct" statistical approach (see corresponding section in Materials and Methods).

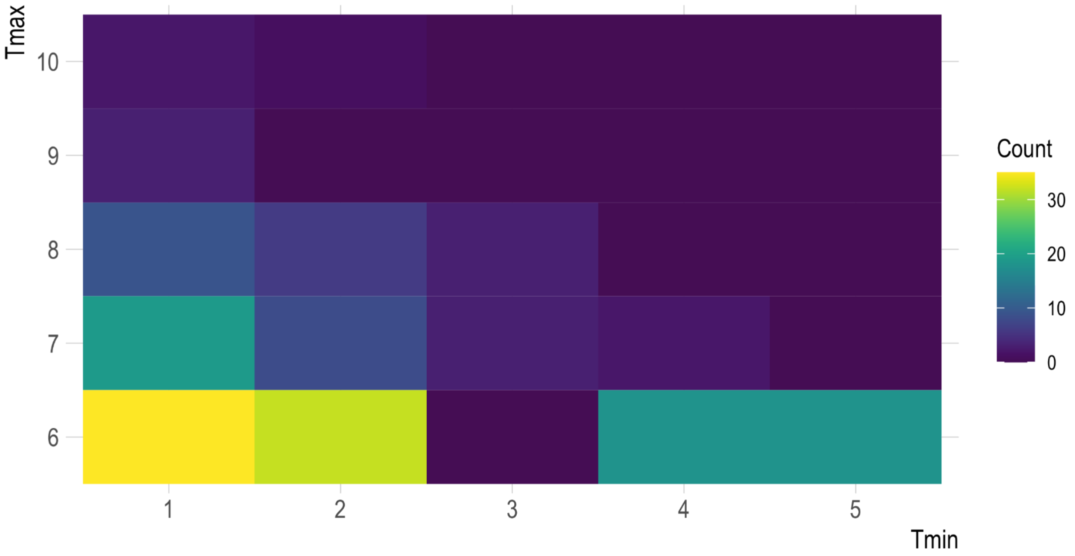
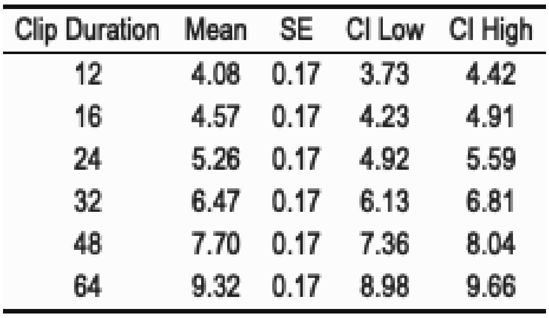


Figure S2. Significant instances across threshold parameters. Heatmap representing the number of significant occurrences extracted from the Direct approach when re-fitting our computational model across different combinations of threshold parameters (Tmin and Tmax).

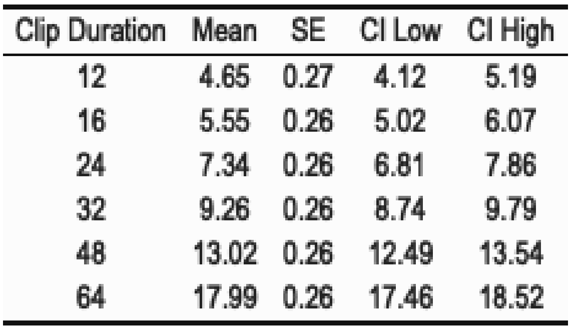
Event Annotations by Methods - Marginal Means and Contrast Analysis

As mentioned in Results, here we report the tables summarising marginal means of event boundaries estimated at different levels of clip duration by linear mixed-effect models fitted on human, GSBS and model-based annotations. In addition, we also report the summary table of the contrast analysis that we run across different scene type levels for human and model-based raters.



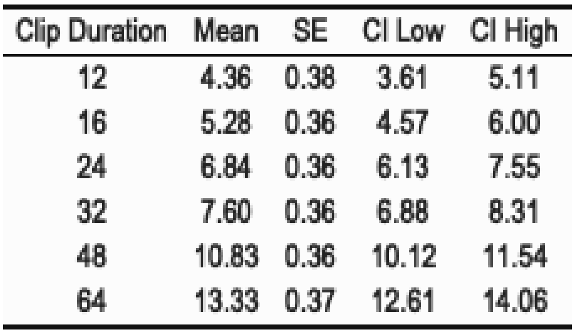
Clip Duration	Mean	SE	CI Low	CI High
12	4.08	0.17	3.73	4.42
16	4.57	0.17	4.23	4.91
24	5.26	0.17	4.92	5.59
32	6.47	0.17	6.13	6.81
48	7.70	0.17	7.36	8.04
64	9.32	0.17	8.98	9.66

Table S1. Human raters: Mean number of estimated boundaries across Clip Durations. Table including the mean number of estimated event boundaries by the linear mixed-effect model fitted on GSBS annotations.



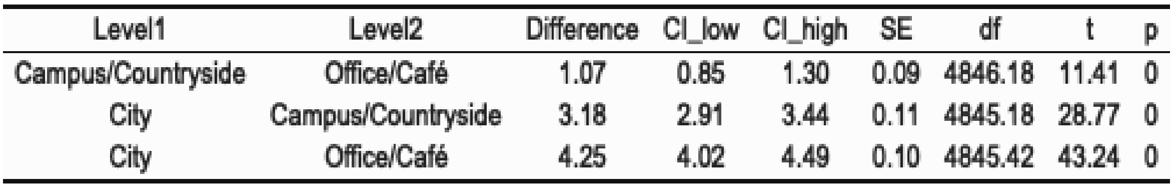
Clip Duration	Mean	SE	CI Low	CI High
12	4.65	0.27	4.12	5.19
16	5.55	0.26	5.02	6.07
24	7.34	0.26	6.81	7.86
32	9.26	0.26	8.74	9.79
48	13.02	0.26	12.49	13.54
64	17.99	0.26	17.46	18.52

Table S2. GSBS: Mean number of estimated boundaries across Clip Durations. Table including the mean number of estimated event boundaries by the linear mixed-effect model fitted on GSBS annotations.



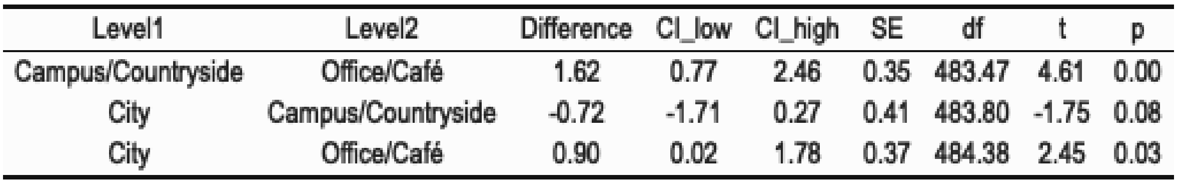
Clip Duration	Mean	SE	CI Low	CI High
12	4.36	0.38	3.61	5.11
16	5.28	0.36	4.57	6.00
24	6.84	0.36	6.13	7.55
32	7.60	0.36	6.88	8.31
48	10.83	0.36	10.12	11.54
64	13.33	0.37	12.61	14.06

Table S3. Model-based: Mean number of estimated boundaries across Clip Durations. Table including the mean number of estimated event boundaries by the linear mixed-effect model fitted on model-based annotations.



Level1	Level2	Difference	CI_low	CI_high	SE	df	t	p
Campus/Countryside	Office/Café	1.07	0.85	1.30	0.09	4846.18	11.41	0
City	Campus/Countryside	3.18	2.91	3.44	0.11	4845.18	28.77	0
City	Office/Café	4.25	4.02	4.49	0.10	4845.42	43.24	0

Table S4. Human raters: Contrast analysis between different Scene Types. Table including the post-hoc contrasts estimated for different scene types from the linear mixed-effect model fitted on human raters annotations.

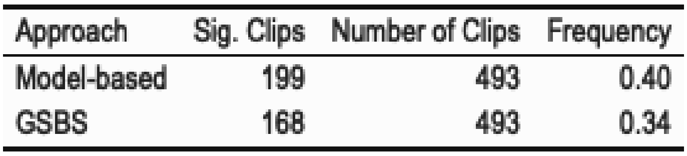


Level1	Level2	Difference	CI_low	CI_high	SE	df	t	p
Campus/Countryside	Office/Café	1.62	0.77	2.46	0.35	483.47	4.61	0.00
City	Campus/Countryside	-0.72	-1.71	0.27	0.41	483.80	-1.75	0.08
City	Office/Café	0.90	0.02	1.78	0.37	484.38	2.45	0.03

Table S5. Model-based: Contrast analysis between different Scene Types. Table including the post-hoc contrasts estimated for different scene types from the linear mixed-effect model fitted on model-based annotations.

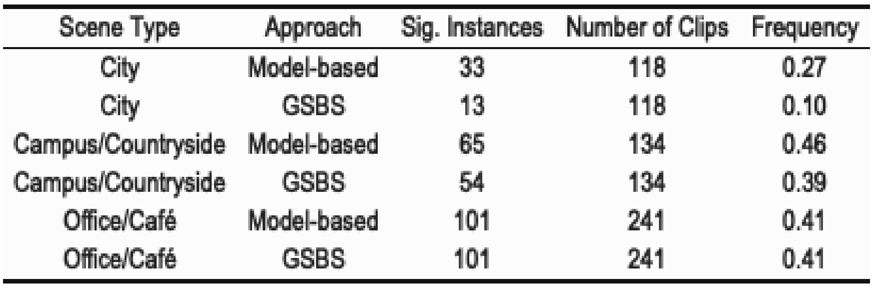
Meta Approach: uncorrected results

While all Results related to our Meta approach were reported after correction for multiple comparison, here we report the summary of the uncorrected ones.



Approach	Sig. Clips	Number of Clips	Frequency
Model-based	199	493	0.40
GSBS	168	493	0.34

Table S6. Number of Significant Clips (uncorrected). Table including the number and proportion of significant clips (e.g. non matching the same segmentation provided by human raters) computed by our Meta Approach for both our model-based approach and GSBS.



Scene Type	Approach	Sig. Instances	Number of Clips	Frequency
City	Model-based	33	118	0.27
City	GSBS	13	118	0.10
Campus/Countryside	Model-based	65	134	0.46
Campus/Countryside	GSBS	54	134	0.39
Office/Café	Model-based	101	241	0.41
Office/Café	GSBS	101	241	0.41

Table S7. Number of Significant Clips across Scene Types (uncorrected). Table including the number and proportion of significant clips (e.g. non matching the same segmentation provided by human raters) computed by our Meta Approach for both our model-based approach and GSBS across scene types.