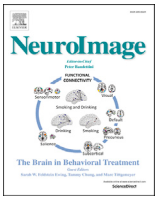
Contents lists available at ScienceDirect

NeuroImage

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



What is changing when: Decoding visual information in movies from human intracranial recordings

Leyla Isik[a,b,*], Jedediah Singer[a], Joseph R. Madsen[c], Nancy Kanwisher[b], Gabriel Kreiman[a]

[a] *Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States*
[b] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States*
[c] *Department of Neurosurgery, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States*

ARTICLE INFO

Keywords:
Electrocorticography (ECoG)
Neural decoding
Object recognition
Ventral pathway
Natural vision
Movies

ABSTRACT

The majority of visual recognition studies have focused on the neural responses to repeated presentations of static stimuli with abrupt and well-defined onset and offset times. In contrast, natural vision involves unique renderings of visual inputs that are continuously changing without explicitly defined temporal transitions. Here we considered commercial movies as a coarse proxy to natural vision. We recorded intracranial field potential signals from 1,284 electrodes implanted in 15 patients with epilepsy while the subjects passively viewed commercial movies. We could rapidly detect large changes in the visual inputs within approximately 100 ms of their occurrence, using exclusively field potential signals from ventral visual cortical areas including the inferior temporal gyrus and inferior occipital gyrus. Furthermore, we could decode the content of those visual changes even in a single movie presentation, generalizing across the wide range of transformations present in a movie. These results present a methodological framework for studying cognition during dynamic and natural vision.

1. Introduction

How does the brain interpret complex and dynamic inputs under natural viewing conditions? The majority of studies in visual recognition have simplified this question by examining neural responses to isolated shapes, presented in static images, with well-defined onset and offset times, and reporting averaged neural signals across multiple repetitions of identical stimuli. The extent to which the principles learned from these studies generalize to the complexities of temporally segmenting and interpreting the kind of rich, dynamic information present in real-world vision remains unclear (Felsen and Dan, 2005; Rust and Movshon, 2005).

Studies of the neural responses to flashing static stimuli along the ventral visual stream have revealed a cascade of computational steps that show progressively increasing shape selectivity and invariance to stimulus transformations (for reviews, see (Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 1999; Connor et al., 2007; DiCarlo et al., 2012)). The starting point to analyze the responses to flashed stimuli involves aligning the neural signals to the stimulus onset, and showing raster plots and post-stimulus time histograms aligned to the transition from a blank screen to a screen containing the stimulus. Despite the trial-to-trial variability in the neural responses elicited by repeated presentation of the same stimulus, several studies have demonstrated that it is possible to read out information about image content in single trials by applying machine learning techniques (reviewed in (Kriegeskorte and Kreiman, 2011)). Furthermore, it is also possible to identify the time at which the stimulus onset happens purely from the neural responses (Hung et al., 2005).

In stark contrast to experiments that present stimuli with well-defined onsets and offsets, natural viewing conditions require interpreting the visual world from a continuous stream of visual input. These conditions present a series of important challenges: (i) there is no obvious “stimulus onset” to align responses to; (ii) the visual system is continuously bombarded with rapidly changing input; and (iii) natural images are significantly more complex and cluttered than those used in many studies with single shapes on a uniform background. In an attempt to begin to examine how the visual system responds under more naturalistic and dynamic conditions, there has been growing interest in using movies as stimuli in neurophysiological studies (e.g. (Vinje and Gallant, 2000; Lewen et al., 2001; Fiser et al., 2004; Lei et al., 2004; Montemurro et al., 2008; Honey et al., 2012; McMahon et al., 2015; Podvalny et al., 2016)) and also in non-invasive studies (e.g. (Hasson et al., 2004; Bartels and Zeki, 2005; Whittingstall et al., 2010; Nishimoto et al., 2011; Huth et al., 2012; Conroy et al., 2013; Russ and Leopold, 2015)).

These studies have demonstrated that general principles of visual

[*] Corresponding author. Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States.
E-mail address: lisik@mit.edu (L. Isik).

http://dx.doi.org/10.1016/j.neuroimage.2017.08.027
Received 28 April 2017; Received in revised form 7 August 2017; Accepted 8 August 2017
Available online xxxx

processing derived from flashing static stimuli are maintained when considering dynamic stimuli but they have also highlighted important differences. For example, in primary visual cortex, investigators have reported that models built from responses to flashed gratings fail to capture all the variance in the neural responses to movies (Vinje and Gallant, 2000; Carandini et al., 2005). In higher visual areas, the responses to complex shapes such as faces are strongly modulated by the dynamic aspects of movie stimuli (McMahon et al., 2015; Russ and Leopold, 2015).

Here we describe a methodology to examine neural responses obtained from intracranial field potentials (IFP) in human epilepsy patients while they passively watched commercial movies. We tackled the central questions defined above by directly using the neural signals to: (i) evaluate *when* there are large visual changes in the continuous visual inputs, and hence how to align neural signals in response to movies, and (ii) identify *what* is the content in the changing movie frames, despite the complex and heterogeneous variations in the movies. In a first experiment, we presented multiple repetitions of short movie clips. We showed that we could decode intracranial field potentials to determine when a visual change happened and identify what changed in those visual events, generalizing across the transformations present in movie clips. In a second experiment, we extended this methodology to the analysis of neural responses to single presentations of a full-length movie.

2. Material and methods

Raw data and code for this manuscript are available at http://klab.tch.harvard.edu/resources/Isiketal_whatchangeswhen.html.

2.1. Physiology subjects

Subjects were 15 patients (ages 4–36, 8 males, 2 left handed) with pharmacologically intractable epilepsy treated at Children's Hospital Boston (CHB) or Brigham and Women's Hospital (BWH). They were implanted with intracranial electrodes to localize seizure foci for potential surgical resection (Ojemann, 1997; Liu et al., 2009). All studies described here were approved by each hospital's institutional review board and were carried out with the subjects' informed consent. Electrode types, numbers and locations were driven solely by clinical considerations.

2.2. Recordings and data preprocessing

Subjects were implanted with 2 mm diameter intracranial subdural electrodes (Ad-Tech, Racine, WI, USA) that were arranged into grids or strips with 1 cm separation. Each subject had between 26 and 144 electrodes (86 ± 26, mean ± SD). We conducted two experiments (described below). We studied a total of 1 284 electrodes (954 in Experiment I, and 330 in Experiment II, Supplemental Table 1 and Supplemental Table 2). All data were collected during periods without seizures or immediately following a seizure. Data were recorded using XLTEK (Oakville, ON, Canada) and BioLogic (Knoxville, TN, USA) with sampling rates of 256 Hz, 500 Hz, 1 000 Hz or 2000 Hz.

For each electrode, a notch filter was applied at 60 Hz and harmonics, and the common average reference computed from all channels was subtracted. We focused on the broadband voltage signals in the 0.1–100 Hz range (referred to as broadband signals throughout the manuscript). In the Supplementary Material, we also considered the power in the intracranial field potential signals filtered in the following broadband frequency ranges: alpha (8–15 Hz, alpha broadband), low gamma (25–70 Hz, low gamma broadband), and high gamma (70–120 Hz, high gamma broadband). All of these are broadband frequency ranges and not single frequency oscillatory signals. After notch filtering, signals were band passed filtered in each of those frequency bands. Power in each frequency band was extracted using a moving window multi-taper Fourier transform (Bokil et al., 2010) with a time-bandwidth product of five tapers. The window size was 200 ms with 10 ms increments.

2.3. Electrode localization

Electrodes were localized by coregistering the preoperative MRI with the postoperative computerized tomography (CT) (Liu et al., 2009; Destrieux et al., 2010; Tang et al., 2014). For each subject, the surface of the brain was reconstructed from the MRI and then assigned to one of 75 anatomically defined regions by Freesurfer. Each surface was also coregistered to a common brain (Freesurfer fsaverage template) for display purposes only, all analyses separating electrodes by brain region were based on localization in individual subject's own anatomical images. We emphasize that all electrode locations are strictly dictated by clinical criteria. In this type of study, comparisons across subjects are complicated because not all subjects have electrodes in the same anatomically defined brain region and there are also differences in electrode locations within each such region across subjects. The locations of the electrodes in Experiment I are shown in Fig. 4A, and the locations of the electrodes in Experiment II are shown in Fig. 7A. Tables S4–S5 report the number of subjects contributing to each anatomically defined brain region in experiment I and II, respectively.

2.4. Neurophysiology experiments

2.4.1. Experiment I

In the first experiment, 11 subjects viewed three 12 s cartoon clips from two separate movies (example frames for one of these movies are shown in Fig. 1A). Each clip was repeated multiple times, between 10 and 68 repetitions (see Supplemental Table 1), depending on subject fatigue and clinical constraints. Clips were presented in a random order with a 1 s interval between clips. Subjects passively viewed the clips. Clips were presented at approximately 4 × 3 degrees of visual angle. Clips were shown in color and had no sound.

2.4.2. Experiment II

In the second experiment, 4 different subjects viewed a full-length commercial movie: Home Alone (subject 12, see example frames in Fig. 1B), Charlie and the Chocolate Factory (subjects 13–14) or In the Shadow of the Moon (subject 15). Movies were presented with sound and color at ~18 × 12 degrees of visual angle. Subjects passively viewed the movies once through. The movies were interleaved with static images presented for a separate experiment. The movie was played for 25s, followed by 20 static images from different categories, and then again by the next 25s of movie.

2.5. Eye tracking experiment

Even though the stimulus size was relatively small to prevent large eye movements, we performed a post-hoc experiment to evaluate whether subjects generate consistent saccades under these viewing conditions (consistency within a subject across repetitions of the same clip and consistency across subjects). A post-hoc eye tracking experiment was conducted on 7 in-lab subjects to examine eye movements. Each subject viewed each 12s clip in Experiment 1, presented five times in a random order. The viewing conditions were the same as in the physiology experiments. Eye position was recorded with an infrared camera eye tracker (EyeLink D1000, SR Research). The median eye position across subjects and repetitions is shown in Fig. S1.

2.6. Data analyses

2.6.1. Cut detection

Movies were segmented based on sharp visual transitions between scenes referred to as movie cuts throughout (see examples in Fig. 1). In Experiment I, the cuts within the 12s clips were manually labeled. In

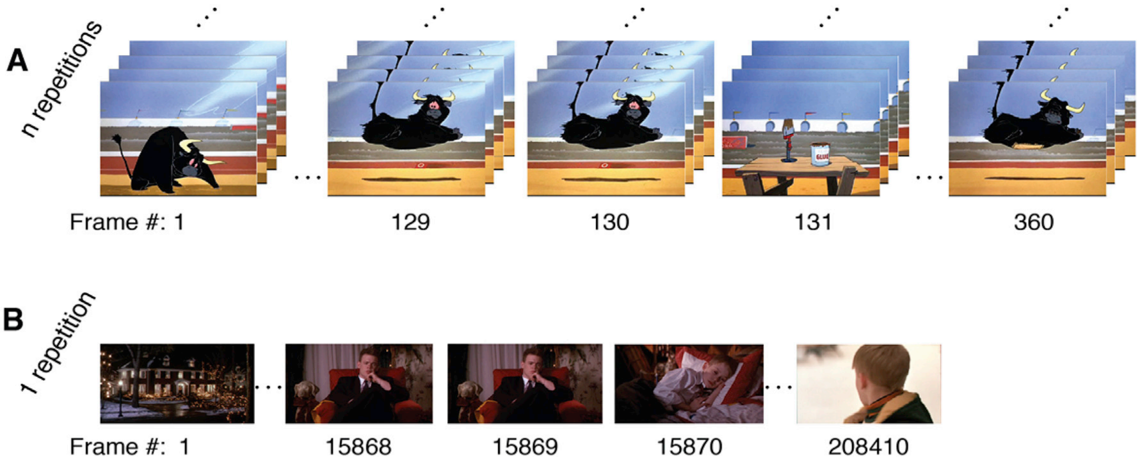


Fig. 1. Experimental paradigm and movie cuts. **A**. Experiment I - Three 12s clips from commercial cartoon movies were presented multiple times without sound, at 30 frames per second, and subtending ~4 × 3 degrees of visual angle (see Supplemental Table 1). The first frame, three middle frames (demonstrating a movie cut between frames 130–131), and the final frame from clip 1 are shown (Methods). Subjects passively viewed the 12s movie clips. **B**. Experiment II – A full-length movie was shown once through with sound, at 24 frames per second, and subtending ~18 × 12 degrees of visual angle. We considered data from patients watching one of three movies in this study (Methods). Example frames from one of the movies, Home Alone 2, including the first frame, three middle frames (demonstrating a movie cut between frames 15,869–15870), and the final frame are shown. Subjects passively watched these full-length movies.

Experiment II, the cuts in the full-length movies were first detected automatically using an algorithm calculating and thresholding pixel differences between consecutive movie frames. The automatically detected movie cuts were then checked and refined manually. We refer to a “shot” as the time period in between two adjacent cuts and we refer to an “event” as a single occurrence of a shot.

2.6.2. Movie labeling

We manually labeled shots in the movies by assigning one label to an entire segment between movie cuts (shots ranged in length from 0.4s to 3.73 s, with an average length of 1.67 s). The objects and background within a given shot are generally different than those in the previous shot and are approximately constant throughout a shot.

In Experiment I, we labeled the presence or absence of the main characters (humanized versions of cartoon animals) in each 12 s clip. This allowed us to test visual selectivity for each repeated event (e.g. appearance of a particular shape) across the course of the movie. In particular, in Experiment I, both 12 s clips contained shots with a single animal, and shots with no animal. Two pairs of animal/no-animal scenes were selected in each 12 s clip, one pair occurring at the beginning of the clip and one pair at the end. In the decoding analyses described below, pairs that were close in time were selected as foils (e.g. each animal shot was closer in time to its no-animal foil than to the other animal shot) so that the decoding algorithm could not simply exploit correlations in the physiological data that occur due to temporal proximity.

In Experiment II, we labeled in each movie shots with a single face and shots with no faces or bodies. Faces were selected as a target for visual decoding because they are a consistent, repeating visual element in all movies shown.

2.6.3. Correlation analyses

In Experiment I, we evaluated how consistent the neural signals were across the repeated presentation of the same 12 s clip for all the cut-responsive electrodes. We correlated the time courses across repetitions of the same 12 s clip. For each of the $n = 954$ electrodes, we calculated the Pearson correlation coefficient between each pair of repetitions in every 50 ms overlapping bin (step size of 1 sample) in each of the three 12 s clips (correlations for an example electrode during one movie clip are shown in Fig. 2D). The choice of a 50 ms window was dictated by the attempt to make the window as small as possible while keeping a sufficient number of sampled voltage values to compute a correlation. To quantify the statistical significance of the correlation coefficients thus obtained, we defined a null distribution by computing the correlation coefficients between each 50 ms bin in the movie and random temporally offset segments. We defined a segment as showing a significant consistency across repetitions when the correlations between repetitions were significantly above chance in at least 20 consecutive 50 ms bins with $p < 0.01$ with respect to the null distribution (e.g. horizontal marks in Fig. 2D). To examine how the timing of consistent responses across repetitions revealed by the inter-repetition correlations related to movie cuts, we calculated the latency between the onset of significantly above chance consistency segments and the previous movie cut (Fig. 3).

We repeated the above correlation analyses using a binning window of 400 ms in Figs. S13B, E, H. This longer time window implies more time points in the calculation of each correlation coefficient. To ensure that this increase in the number of time points would not bias the results, we repeated the analyses with a bin size of 400 ms and a smoothing factor of 8 in Fig. S13C, F and I to match the number of time points in Fig. 3. Given the larger time window in the analyses in Fig. S13, we explicitly removed windows that intersected a camera cut (to avoid, for example, a window from −200 to +200 ms with respect to a movie cut to be assigned to −200 ms and erroneously suggest windows of high correlation before movie cuts).

2.6.4. Cut responsiveness

To examine whether the physiological signals showed a significant evoked response to cuts (e.g. Fig. 2B), we compared the IFP response, defined as the range (max-min) of the broadband signals or the total power in each frequency band in the 50–400 ms post-cut window to the corresponding values in the −400 to −50 ms pre-cut window. We defined cut responsive electrodes as those that showed a $p < 0.01$ difference in the post-cut versus the pre-cut windows when considering all repetitions of the $n = 20$ cuts (all cuts, excluding the first cut – i.e. movie onset – in each movie) based on a permutation test where the pre-cut and post-cut windows were randomly shuffled 1 000 times to define a null distribution. Channels that yielded a greater IFP response than 99% of the null distribution were defined as significant with $p < 0.01$. All of the electrodes that met this significance criterion are reported in Supplemental Table 2 through 5 and in Section 3.1.

2.6.5. Decoding methods

Several analyses in the manuscript describe the accuracy in discriminating between visual events during the movie using statistical classifiers. We describe next the methods for those analyses.

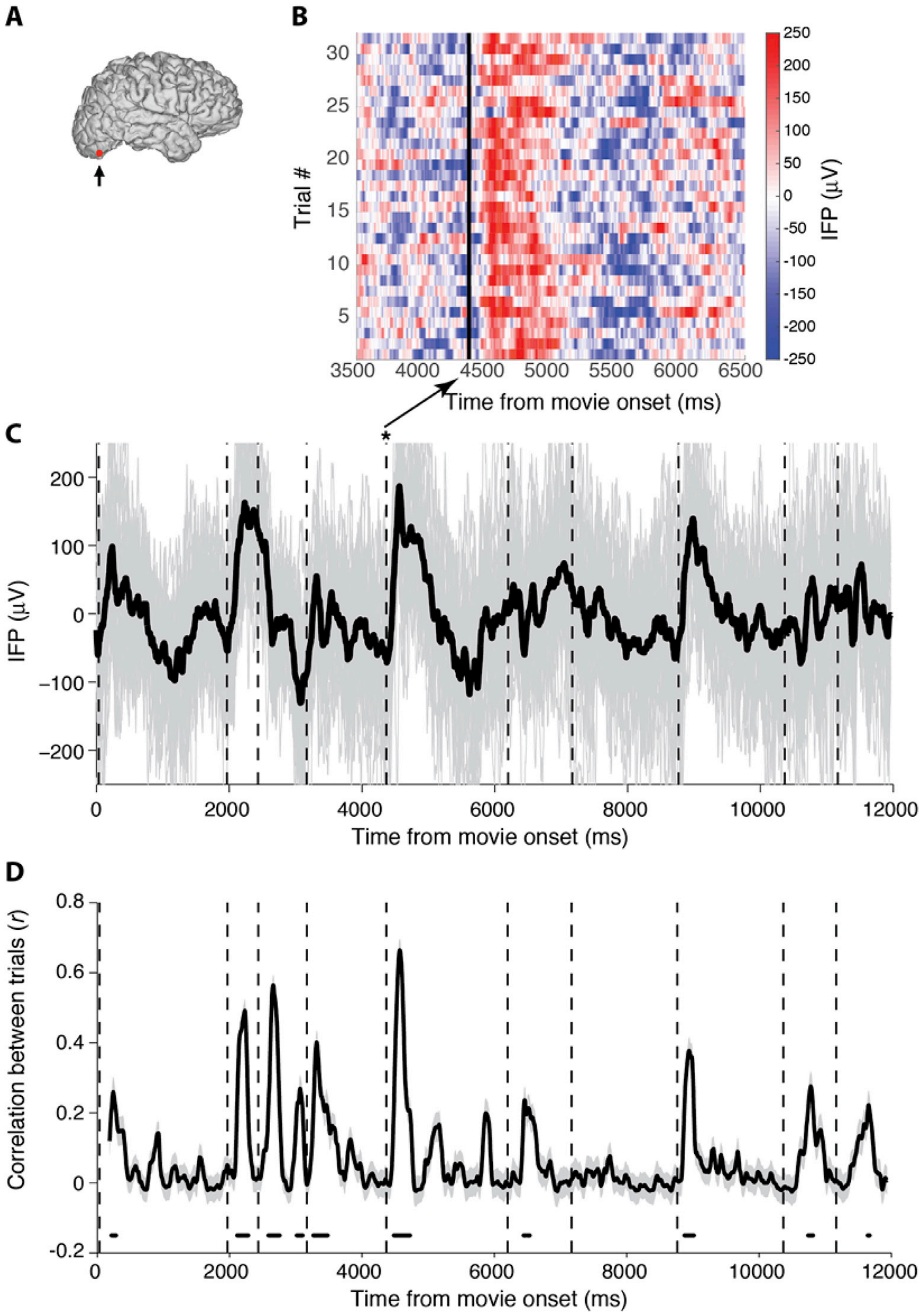
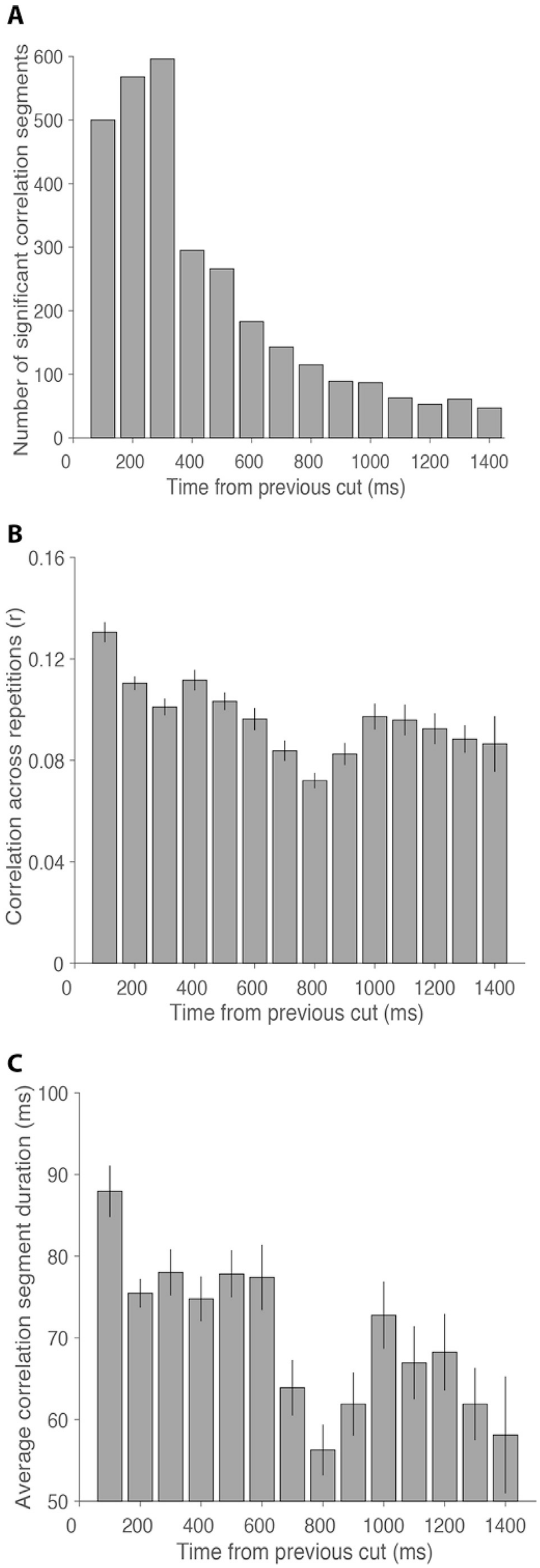


Fig. 2. Example electrode showing consistent physiological responses to movie cuts (Experiment I). **A**. Electrode location. The electrode was located in the right inferior occipital gyrus (Talairach coordinates = [35.9, −82.8, −14.5]). **B**. Raster plot showing the intracranial field potential (IFP) surrounding the cut transition shown in Fig. 1A (frame 130–131 in movie clip 1). Each row denotes a repetition of the movie (n = 32 repetitions). The color indicates the IFP at each time point (bin size = 3.9 ms, see color scale on right). The movie cut triggered a large change in voltage in almost every repetition. **C**. Electrode's broadband voltage time course over the entire 12 s movie clip 1. Mean activity is shown with a thick black line, and 32 individual repetitions are shown with gray traces. Dashed vertical lines indicate movie cuts, and the cut shown in **B** is indicated with an asterisk. Several, but not all, of the cut transitions elicited large voltage changes that can be observed even in individual repetitions. The y-axis is cut at −250 and 250 μV but some individual traces extend beyond these limits. **D**. Average pairwise correlation (Pearson coefficient, *r*, mean ± SEM) across the 32 choose 2 (496) pairwise comparisons between repetitions calculated in 50 ms non-overlapping bins. Horizontal black lines at the bottom of the plot indicate time periods when the average pairwise correlation across repetitions was significantly above chance based on a $p < 0.01$ permutation test (Methods). See Fig. S2 for a similar example in the high gamma frequency band.

2.6.5.1. Classifier features. In each decoding analysis, we considered the average voltage in 50 ms non-overlapping time bins for each electrode as input to the classifiers described below. In the Supplementary Material we repeated these analyses examining the average power in the alpha (8–15 Hz), low gamma (25–70 Hz), or high gamma (70–120 Hz) frequency ranges. Depending on the specific analyses, we used either single electrodes, pseudo-populations composed of a fixed number of electrodes per region or a population from multiple electrodes selected across subjects, as described below. The entire decoding procedure was repeated in each 50 ms bin.



In Experiment I, because subjects viewed multiple repetitions of identical stimuli, electrodes were pooled across all subjects into pseudo-populations for specific locations. We first examined the decoding performance in each brain region by pooling electrodes within a given anatomical parcel from the Freesurfer Destrieux atlas (Section 2.3). For this analysis, we considered all anatomical parcels with at least 8 electrodes, and performed decoding with the pattern of activity across the top 8 electrodes (as measured by the electrode selection procedure described below) in each of these regions (Fig. 4B–C, Fig. 5B–C). Next, we also evaluated performance by combining electrodes across separate brain regions and subjects (Fig. 5D (Tang et al., 2014)).

In Experiment II, because subjects did not all view the same movie, decoding was performed separately for each electrode and subject. We calculated decoding performance per brain region with at least 5 electrodes by averaging the single electrode decoding results for all electrodes in each anatomical region (Fig. 7B–C). We also pooled all electrodes per subject and movie to perform population level decoding, and then again averaged the decoding results post-hoc across subjects (Fig. 7D).

2.6.5.2. Feature pre-processing. The data from each electrode (feature) was z-scored normalized based on the mean and standard deviation in the training data. In addition, an ANOVA was performed on each input feature using only the training data. The ANOVA selects electrodes that show a larger variance between “categories” than within a “category” as described next. In Figs. 4B and 7B, the ANOVA analysis was used to select those electrodes that showed a larger variance between cuts and non-cuts compared to the variance within repetition of cuts. In Fig. 5B, the ANOVA was used to select electrodes that showed a larger variance between different movie shots compared to the variance within the same movie shots. In Fig. 5C–D, the ANOVA was used to select electrodes that showed a larger variance between shots with an animal and shots without an animal compared to the variance within shots with an animal and within shots without an animal. This method has been shown empirically to improve the signal to noise ratio of decoding with human MEG and monkey LFP time series data (Meyers et al., 2008; Isik et al., 2014).

2.6.5.3. Classifier. Decoding analyses were performed using a maximum correlation coefficient classifier. This classifier computes the correlation between each test data point and the mean of all training data points from each class. Each test point is assigned the label of the class of the training data with which it is maximally correlated (Fig. S12A).

2.6.5.4. Cross-validation. For each decoding run, the data were divided into 10 cross-validation splits. Feature pre-processing (z-scoring and ANOVA) was performed on 9 out of 10 of the splits, and testing was performed on the 10th held out split.

The decoding at each time bin was repeated for 20 times, each with a different train/test data split. The average performance of the 20 decoding runs is displayed as “classification accuracy” as a function of time from cut onset in Figs. 5D and 7D. In other cases, we summarized classification accuracy by reporting the average value from 50 to 400 ms post-cut onset (Figs. 4B and 5B-C, 7B-C).

Fig. 3. Properties of neural responses that were consistent across trials. **A.** Distribution of the onset of segments with statistically significant correlation across repetitions in all electrodes (n = 954), calculated with a sliding window of 50 ms duration, as a function of time from the previous cut. Bin size = 100 ms (**Methods**). These segments of consistent correlation across repetitions begin mostly within the 300 ms following a cut. **B.** Average correlation coefficient between repetitions in each time bin for all the segments with statistically significant correlation between repetitions in **A** (mean ± SEM). **C.** Average duration between the beginning of the first and last time points for all the consecutive segments with statistically significant correlation between repetitions in **A** (mean ± SEM). See Fig. S3 for corresponding analyses in different frequency bands and Fig. S13 for the same analyses using different window sizes.

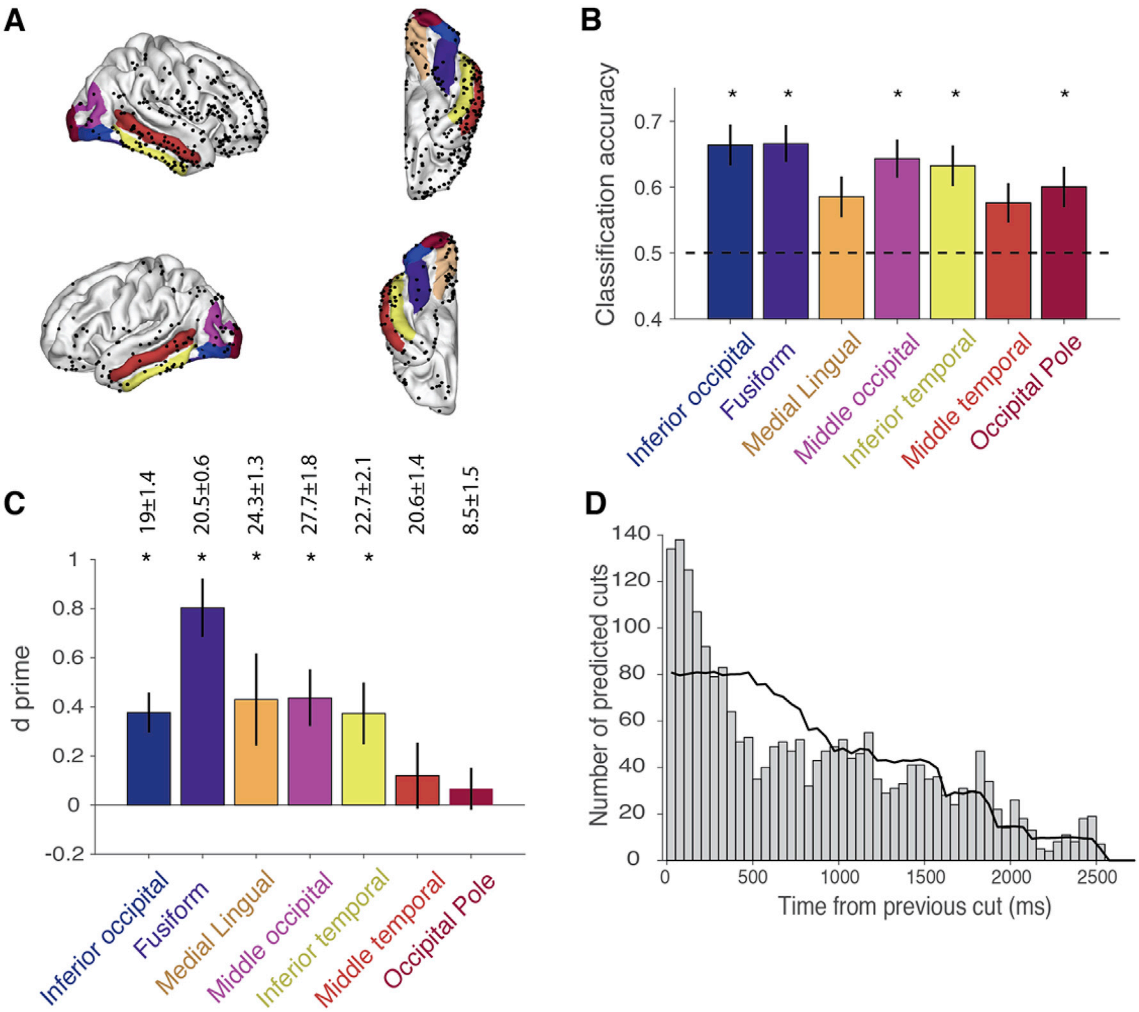


Fig. 4. Movie cuts and shots can be decoded from ventral visual cortex regions. **A**. Location of all electrodes in Experiment I projected onto a common reference brain (Freesurfer fsaverage brain) shown at lateral and ventral views. Each dot corresponds to one electrode (total = 994 electrodes, Supplemental Table 1). Seven anatomical regions (out of 25 regions with at least eight electrodes) with significantly above chance decoding performance in any of the decoding tasks in **B** or **C** are highlighted. **B**. Classification accuracy from n = 8 electrodes in each region, between movie segments with a cut versus segments without a cut in the seven regions highlighted in Fig. 4A (mean ± SD across 20 decoding runs, Methods). Chance = 0.5. The classification accuracy is reported as the average from 50 to 400 ms post cut onset. Asterisks indicate significant decoding based on a $p < 0.01$ permutation test (Methods, Supplemental Fig. 12A). **C.** Sensitivity (d') to detect visual transitions during the entire 12 s clip time course for held out repetitions of movie clips 1 and 2 (mean ± SD across 20 decoding runs, **Methods**, Supplemental Fig. 12B). Number at the top of each bar plot indicates the number of predicted visual transitions per region (the actual number of all cuts in movie clips 1 and 2 was 17). **D.** The bars show the latency difference between the time of the predicted visual transitions (first time point in visual transition predicted periods, see Fig. S12B) in **C** and the time of the previous true cut for the five regions with significantly above chance d' values in **C**. Bin size = 50 ms. The line shows the average distribution obtained from randomly selecting the same number of times as predicted visual transitions. The distribution of selected transition times is significantly different from the random distribution ($p < 10^{-10}$, Kolmogorov-Smirnov test). See Fig. S7 for corresponding analyses in different frequency bands.

2.6.5.5. Decoding analyses, experiment I.

(i) We compared movie segments with a movie cut versus random segments falling at least 400 ms away from a movie cut (Fig. 4B).

(ii) We evaluated whether we could detect visual transitions in the entire 12 s clip. The procedure is illustrated in Fig. S12B. We used the average vector representing "cut" and "no-cut" events as described in (i) and Fig. S12A. For each 50 ms window from held-out repetitions, if the correlation with the "cut" vector was larger than the correlation with the "no-cut", we assigned a label of +1, otherwise we assigned a label of −1. We defined hits as those 50 ms windows which had a label of +1 and which were within the 0–400 ms after a cut. Similarly, we defined false alarms as those 50 ms windows which had a label of +1 and which did not occur within 0–400 ms after a cut. We calculated the d prime measure across all 50 ms time bins in the 12s clip: d prime = Z(hit rate) – Z(false alarm rate), where Z is the inverse cumulative distribution function (Fig. S12B, Fig. 4C). We defined a predicted visual transition as a set of 1 or more continuous 50 ms windows classified as +1. For each predicted visual transition, we defined the time of the transition as the first 50 ms window in the set. We calculated how far away those predicted visual transitions were from the nearest prior cut in Fig. 4D.

(iii) We tested for visually selective signals by decoding the different camera shots from each other (Fig. 5B). We included the 13 camera shots in the first two movies (all the movie cuts that were presented at least 20 times across subjects, see Supplemental Table 1, Fig. S6, excluding the first and last shot).

(iv) We compared shots with an animal versus shots without an animal (Fig. 5A,C-D). We performed this animal versus no animal decoding first across repetitions of the same movie clips (referred to as the "within shot" condition). Next, we decoded across shots in the same 12 s clips (referred to as the "across shot" condition),

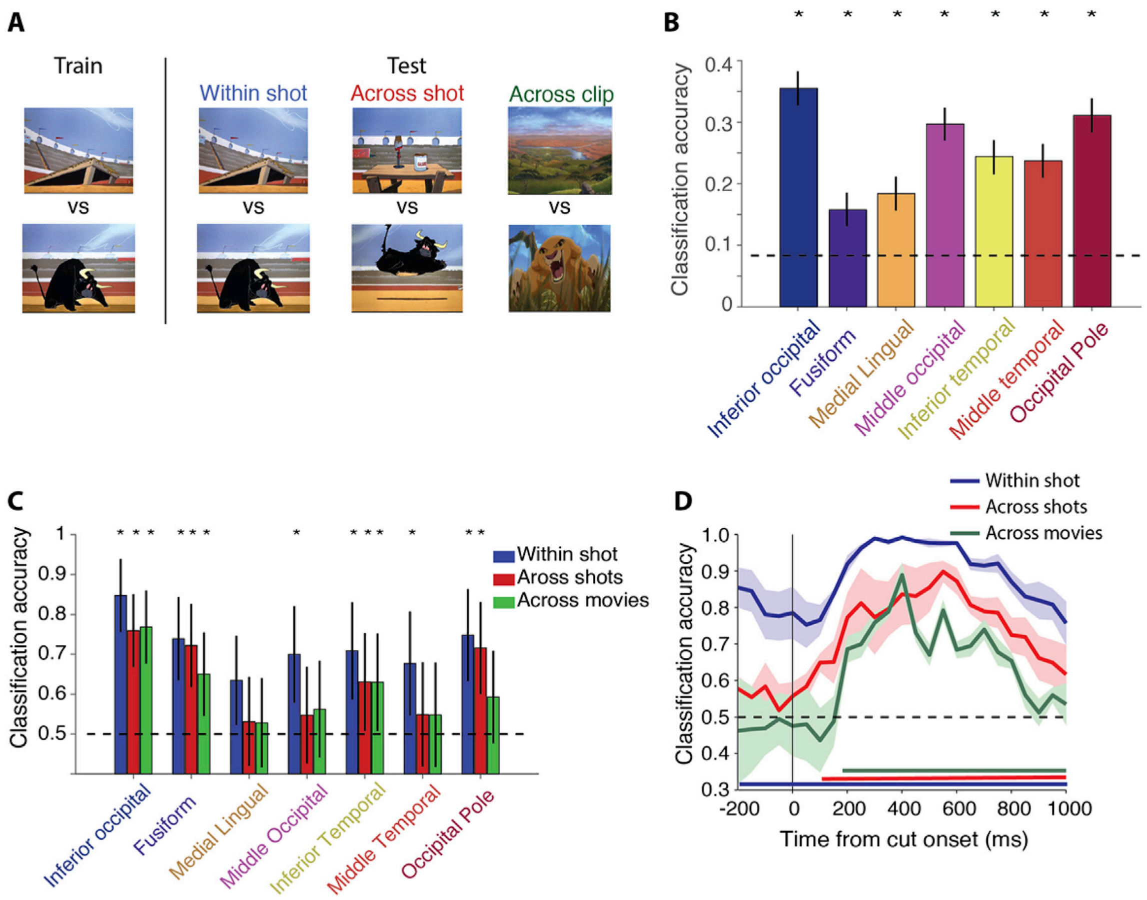


Fig. 5. Visual information generalizes across movies in 12s clips. **A**. We decoded shots with an animal versus shots without an animal, first from repetitions of the exact same shot pairs ("within shot", blue), next with generalization across different shots in the same movie clip ("across shot", red), and finally across movie clips ("across clips", green). One example pair of frames (first frame in shot) depicting the different conditions is shown. Decoding was repeated for four pairs of clips (from two of the three 12s clips that represent the two unique movies, Fig. S4, Methods). **B**. Classification accuracy to label each of the 13 cuts from clips 1 and 2 (excluding the first and last cut from each movie, Fig. S6) using n = 8 electrodes in each of the seven regions highlighted in **A** (mean ± SD across 20 decoding runs, Methods). Chance = 1/13. The classification accuracy is reported as the average from 50 to 400 ms post cut onset. Asterisks indicate significant decoding based on a $p < 0.01$ permutation test (Methods). See Fig. S7 for corresponding analyses in different frequency bands. **C**. Classification accuracy from n = 8 electrodes in each region for shots with versus without an animal (mean ± SD across 20 decoding runs, chance = 0.5) in the seven highlighted regions described in Fig. 4A. We considered 3 conditions corresponding to different levels of extrapolation: within shot (blue), across shots (red), and across movies (green). The classification accuracy is reported as the average from 50 to 400 ms post cut onset. Region labels are color coded following the conventions in Fig. 4A. Asterisks indicate significant decoding for each of the three decoding conditions based on $p < 0.01$ permutation test (Methods). **D**. Visualization of dynamic classification accuracy for shots with an animal versus without an animal across time relative to cut onset from a pseudo population based on feature selection from all electrodes across all subjects (mean ± SD across 20 decoding runs, Methods). Feature selection was applied at each time point to choose selective electrodes in the training data to be used in the classifier (Methods). Horizontal line indicates chance classification. Note that the 'within shot' classification accuracy was significantly above chance even before the cut onset, because the visual stimulus pre-cut was identical in the training and test sets (see discussion in text). While the 'Within shot' classification accuracy was significantly above chance for the entire time course, the 'Across shot' and 'Across clip' classification accuracies were significantly above chance from 100 to 1 000 ms and 200–1 000 ms post-cut onset, respectively. See Fig. S9 for corresponding analyses in different frequency bands.

and finally and across shots from different movie clips (referred to as the "across clip" condition).

2.6.5.6. Decoding analyses, experiment II.

(i) We compared movie segments with a movie cut versus random segments falling at least 400 ms away from a movie cut, as in experiment I (Fig. 7B).

(ii)We compared shots with a single face versus shots with no face (Fig. 7C–D).

3. Results

We investigated the neurophysiological responses elicited by dynamic movie stimuli by recording intracranial field potential (IFP) signals from 1 324 electrodes implanted in 15 patients with epilepsy (Tables S1–S3). We conducted two experiments: (i) Experiment I consisted of repeated presentation of three 12s commercial cartoon movie clips (Fig. 1A and 954 electrodes); (ii) Experiment II consisted of a single presentation of a full-length commercial movie (Fig. 1B and 370 electrodes).

3.1. Neurophysiological responses to time-varying stimuli (Experiment I)

In multiple Visual Neuroscience experiments, stimuli are flashed with well-defined onset and offset times and responses are analyzed by aligning activity to the appearance of a stimulus. Movies, as a coarse proxy to natural vision, lack those stimulus onsets. We conjectured, with others (McMahon et al., 2015), that the drastic changes between consecutive frames that occur during movie cuts provide a strong temporal demarcation. Fig. 1 shows two examples of movie cuts (transition from frame 130 to 131 in Fig. 1A and from frame 15,869 to 15,870 in Fig. 1B) and the accompanying large changes in the visual field. We set out to investigate whether such movie cuts trigger the onset of physiological responses and can thus be used to demarcate visual events in movies.

We started by aligning the IFP signals to movie cuts. Fig. 2B shows the responses of an example electrode located in the right inferior occipital gyrus (Fig. 2A) that demonstrated a vigorous modulation after one of the movie cuts. The changes in IFP were evident in almost every single repetition of the movie clip, showed a consistent latency of approximately 100 ms after the cut and were short-lived, with the voltage returning to baseline within approximately 400 ms after the cut. This electrode showed an appreciable modulation elicited by most, but not all, the cuts in the 12 s clips (Fig. 2C). To further quantify the modulation in IFP, we computed the degree of consistency in the responses evaluated by the Pearson correlation coefficient between the voltage time series for every possible pair of repetitions, using a window of 50 ms (Fig. 2D). The correlation coefficient largely hovered around zero, indicating that the IFP signals were inconsistent across repetitions, except for sharp spikes in correlation, which were typically evident right after a movie cut. For the example electrode in Fig. 2 and movie clip 1, there was a significant increase in consistency after 9 of the 10 movie cuts.

We defined an electrode as visually responsive if the range (max-min) of the broadband IFP signals from 50 to 400 ms after a movie cut was significantly different from the range from −400 ms to −50 ms before a movie cut, using all cuts across the 3 movie clips ($p < 0.01$ permutation test, Section 2.6.4, similar criteria have been used in other work, e.g. (Liu et al., 2009)). In the Supplementary Material, we report the results obtained by evaluating modulation in the alpha (8–15 Hz), low gamma (25–70 Hz) and high gamma (70–120 Hz) bands of the IFP signals.

Using these criteria, out of the total of 954 electrodes in Experiment I, we obtained 51 electrodes, which were mostly located in the occipital pole, and inferior and middle occipital gyri and, to a lesser degree, in the fusiform gyrus, medial lingual gyrus, and inferior temporal gyrus (Table S4). In order to avoid potential physiological changes elicited by eye movements, we kept the stimuli relatively small (~4° × 3°) and we restricted the analyses to the initial neurophysiological response between 50 and 400 ms. Furthermore, we conducted a separate post-hoc experiment in non-epilepsy subjects to measure eye movements under the same stimulus presentation conditions and we did not observe any consistent eye movements elicited by the movie cuts (Fig. S1). Therefore, it seems more likely that the modulatory changes in the physiological signals were triggered by the large changes in the visual stimulus rather than by large saccadic eye movements. Reliable responses triggered by movie cuts were also evident in other frequency bands, an example in the high gamma band is shown in Fig. S2.

3.2. Responses that were reproducible between repetitions largely clustered shortly after movie cuts

We next sought to evaluate the degree of trial-to-trial reproducibility in the physiological responses across the entire 12 s clip and the whole set of electrodes in our sample. We plotted the statistical significance of the correlation coefficient over the entire 12 s clips in each electrode on the Freesurfer fsaverage template brain (Fig. S5A). Multiple electrodes along the ventral stream showed reliable responses (Table S4, see Figs. S5B–D for the results in other frequency bands). As illustrated for the example electrode in Fig. 2, the increase in correlation between repetitions was largely present in the initial ~300 ms after cut onset. We followed the procedure in Fig. 2D to detect segments with statistically significant correlation between repetitions. The majority of consistent responses fell within ~300 ms of a movie cut (Fig. 3A). Throughout the entire population of electrodes, there was a small number of consistent responses occurring >500 ms away from movie cuts (Fig. 3A). For example, there was a small but statistically significant peak before the 3rd cut and another small non-significant peak before the 5th cut in Fig. 2D. However, the degree of consistency, as quantified by the correlation coefficient between repetitions, showed a small drop with the time from movie cut onset (Fig. 3B). Moreover, the duration of those segments showing consistency between repetitions also showed a small decrease as a function of time from the previous cut (Fig. 3C). To further illustrate this point, we searched in the entire electrode sample for two example electrodes with the most reliable response segments that were more than 400 ms away from a movie cut (Fig. S4). Even though the peaks in Fig. S4 represent the strongest examples, they are still weaker and shorter than those illustrated in Fig. 2D. Similar conclusions were drawn when considering other frequency bands (Fig. S3). The correlation coefficients in Fig. 3 were calculated using a window size of 50 ms; similar conclusions were reached when considering a window size of 400 ms (Fig. S13). In sum, the abundance, strength and duration of consistent responses was largely linked to the occurrence of movie cuts.

3.3. Detecting the presence of movie cuts (Experiment I)

Under natural viewing conditions, in the absence of a blank screen followed by a flashed stimulus, the brain needs to determine *when* there is a visual change and *what* that change consists of. The when and what computations need to take place in single events, without averaging. To evaluate whether the neural signals are able to discriminate the timing of changes in the visual world, we built machine learning classifiers to discriminate between movie segments (350 ms duration) containing a movie cut versus movie segments without a movie cut (Fig. 4B). The control movie segments consisted of random time periods that were at least 400 ms away from a cut. We built pseudopopulations of electrodes in different anatomically defined brain regions that contained at least 8 electrodes by pooling data across all patients (Fig. 4A, Section 2.6.5.1). In each region, we used the 8 most selective electrodes per region (as determined by an ANOVA applied to the training data, see Section 2.6.5.2). We report the classification accuracy, i.e., the proportion of repetitions where the machine learning classifier correctly determined the presence or absence of a movie cut (chance = 0.5). Of the 25 regions with at least 8 electrodes (Table S4), 5 regions showed significantly above chance classification accuracy: inferior occipital gyrus, fusiform gyrus, middle occipital gyrus, inferior temporal gyrus and occipital pole. The average classification accuracy across these 5 regions was 0.62 ± 0.04 (mean ± SD, across regions; see Supplemental Figs. 7A–C for the corresponding classification results using IFP signals filtered in different frequency bands).

Whereas the analysis in Fig. 4B compares 350 ms segments with and without cuts, the brain needs to be able to detect those events in single events and during a continuous stream. Next, we developed a classifier to investigate whether it is possible to detect visual transitions in single events during the entire 12s clips (Methods). The procedure is schematically described in Fig. S12B. This classifier continuously determines whether there is a visual transition, thus making correct detections (hits) as well as false ones (false alarms). We evaluated the accuracy of this continuous prediction by measuring the classifier's sensitivity using d prime. We found that classifiers using data from five of the seven regions described in Fig. 4A (excluding the middle temporal gyrus and the occipital pole) detected visual transitions with above chance precision with an average d' of 0.48 ± 0.18 (mean ± SD across significant regions, Fig. 4C). We estimated the latency of these visual transition predicitons by measuring the time difference to the nearest prior cut; the mean latency was 690 ± 610 ms (mean ± SD, across all significant regions, Fig. 4D). The distribution of these time differences was significantly different from the one expected under the null hypothesis defined by 10,000 runs of randomly selecting the same number of time points per movie as predicted transitions (Fig. 4D, black line, $p < 10^{-10}$, Kolmogorov-Smirnov test; see Supplemental Figure D–I for the corresponding classification results using IFP signals filtered in different frequency bands). In sum, it is possible to detect when the image changes within a continuous stream from the neural responses along the ventral visual stream.

3.4. Decoding visual events in movies (Experiment I)

After detecting when there is a visual transition in the movie, we asked whether it is possible to selectively identify *what* visual event changes occur. To address this question, we assessed whether the neural signals could discriminate among the 350 ms windows (ranging from 50 to 400 ms) after movie cuts. We selected 13 movie cuts that were presented at least 20 times (Fig. S6, Section 2.6.5.5). Of the 25 regions with at least 8 electrodes, we found 7 regions that showed above chance classification accuracy based on a $p < 0.01$ permutation test (Fig. 5B). These 7 regions included the 5 regions described in Fig. 4B and also the medial lingual gyrus and middle temporal gyrus. The average classification accuracy across these 7 regions was 0.27 ± 0.07 (chance = 1/13 = 0.08; see Supplemental Figs. 8A–C for the corresponding classification results using IFP signals filtered in different frequency bands).

The results in Fig. 5B show classification accuracy averaged from 50 to 400 ms with respect to movie cuts. To summarize and visualize dynamic changes in classification accuracy as a function of time, we pooled electrodes across all subjects and selected those electrodes that showed larger variation across the 13 movie cuts than within repetitions of the same movie cut using only training data (described under feature selection in **Methods**). We performed the same 13-way cut classification analysis described in Fig. 5B. This analysis shows that classification accuracy started to increase at around 100 ms after a movie cut and peaked at around 400 ms (Fig. S8D), consistent with the example electrode dynamics shown in Fig. 2 and also with previous work decoding different objects with static images (Liu et al., 2009; Tang et al., 2014). Fig. S8D shows that classification accuracy was also high at t = 0, and even *before* the onset of the movie cut. Unlike experiments where static images are presented in random order and are preceded by a blank screen, in the movie presentation, the visual stimulus preceding a movie cut was always the same across different repetitions. Furthermore, several movie cuts were preceded by another movie cut within a few hundred ms (e.g. cut numbers 2 and 3 in movie clip 1, Fig. S6), contributing to the significant classification accuracy before and at t = 0 in Fig. S8D.

3.5. *Invariant decoding of visual events in movies (Experiment I)*

A central challenge in visual recognition involves combining selectivity to different shapes with invariance to the myriad transformations in those shapes (Booth and Rolls, 1998; Riesenhuber and Poggio, 1999; Serre et al., 2007; DiCarlo et al., 2012). After identifying when visual transitions occur and what changes during each event, we asked whether these visual shape-selective signals generalize across transformations in the stimuli. To test the degree of invariance in the visual shape-selective responses, we labeled the content of each shot with the presence or absence of a cartoon humanized animal. We selected four animal/no-animal shot pairs (from movies 1 and 2, Fig. S6, Section 2.6.5.5), and used the same methodology described above to determine in each event whether an animal was present or not, with varying amounts of generalization described next (Fig. 5A).

First, the classifier was trained on a subset of the repetitions and tested on the remaining repetitions of the same shots ("within shot", Fig. 5C, blue bars), requiring generalization across different repetitions of identical stimuli (similar to Fig. 5B, here using a subset of the shots for comparison with the next set of analyses and specifically distinguishing shots containing an animal versus shots not containing an animal, chance = 0.5). As expected from the previous analyses, in Fig. 5C we observed significant classification accuracy in 6 of the 7 regions described in Fig. 4A (the medial lingual gyrus did not reach statistical significance in this analysis). The mean within-shot classification accuracy in these 6 regions was 0.74 ± 0.06 (mean ± SD across regions).

Next, we evaluated the degree of generalization across different shots containing an animal within the same movie ("across shots", Fig. 5C, red bars). To avoid conflating tolerance to different shots with correlated activity in time, each animal versus no animal pair was selected to be closer in time to each other than to the second animal versus no-animal pair (i.e., each shot containing an animal was closer in time to its no-animal foil shot than to the other animal containing shot, Fig. S6). This analysis revealed significant classification accuracy in 4 of the 7 regions described in Fig. 4A: inferior occipital gyrus, fusiform gyrus, inferior temporal gyrus and occipital pole. The mean within-shot classification accuracy in these 4 regions was 0.71 ± 0.05.

Finally, we considered the most extreme case of visual generalization by asking whether we could train a classifier to discriminate shots containing an animal or not in one movie and test it on a different movie ("across clip", Fig. 5C, green bars). Three brain regions, inferior occipital gyrus, fusiform gyrus and inferior temporal gyrus, yielded significant classification accuracy with an average performance of 0.68 ± 0.07.

To summarize and visualize the temporal dynamics in classification accuracy, we followed the procedure described in the previous section for Figs. S8D–G and combined electrodes across all subjects in Fig. 5D. The dynamics revealed an increase in classification accuracy commencing around 100 ms post cut onset and peaking around 400 ms post cut onset.

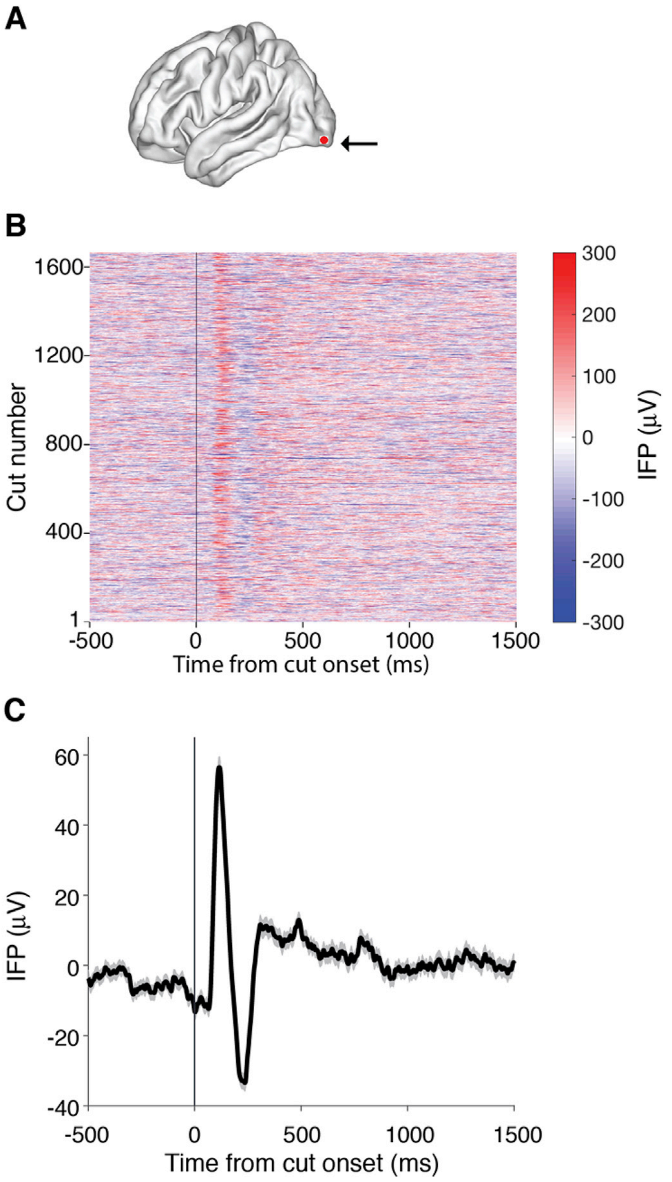


Fig. 6. Example electrode showing a consistent physiological response to movie cuts in full-length movies. **A.** Location of one example cut-responsive electrode (Experiment II) in the left occipital pole (Talairach coordinates = [−2.2, −92.4, −4.3]). **B.** Raster plot showing the intracranial field potential (IFP) surrounding all cut transitions in the full-length movie (Home Alone 2). Each row denotes a different cut (n = 1 630 cuts). The color indicates the IFP at each time point (bin size = 0.5 ms, see color scale on right). **C.** Average IFP time course (mean ± SEM) over all movie cuts. See Fig. S11 for a similar example in the high gamma frequency band.

As noted in Figs. S8D–G, the within-shot condition (blue curve) also revealed strong classification accuracy at and before cut onset in Fig. 5D. Fig. S9 presents corresponding results examining IFP signals filtered in different frequency bands. In sum, the results presented in the previous section and this section show that we can selectively extract information about what changes in the image in single events and with a considerable degree of invariance to the pixel-level transformations.

3.6. Detecting the presence of movie cuts in single presentation of movies (Experiment II)

The insights and analyses derived from Experiment I relied on multiple repeated presentations of the same identical movies. Under natural viewing conditions, the brain must rely strictly on unique presentations of single events. Fig. 5C–D showed that it was possible to decode the presence of absence of an animal by generalizing across different shots and even different movie clips. However, all the classifiers in Fig. 5C–D were still trained using multiple repetitions of identical stimuli. As a more stringent test of generalization across events, we conducted Experiment II where subjects passively viewed a single repetition of a full-length commercial movie. In lieu of identical stimulus repetitions, we leverage the repetition of similar visual events across the duration of a movie.

We assessed whether it was possible to detect when large visual changes occurred in the full-length movies. As described in Fig. 2, neural signals showed strong changes in voltage shortly after movie cuts in the full-length movie. Fig. 6 illustrates the responses of an example electrode located in the left occipital pole that showed consistent (but not identical) changes after almost every movie cut (see raster plot depicting every movie cut in Fig. 6B), despite the fact that the cuts vary enormously in content and were only shown once (see also Fig. S10). The voltage deflections commenced approximately 100 ms after a movie cut (Fig. 6C). In total, we found 61 (out of 330 total) cut-responsive electrodes (Section 2.6.4), located primarily in the cuneus, medial lingual gyrus, fusiform gyrus, inferior occipital gyrus and occipital pole (Table S5).

Following the procedure used in Fig. 4B, we evaluated whether we could distinguish a segment from 50 to 400 ms post cut onset from random time points in single events (Fig. 7B). Because of the smaller total number of electrodes in Experiment II, we considered regions with at least 5 electrodes (as opposed to the threshold of 8 electrodes used in Figs. 4 and 5). Also, because subjects watched different full-length movies, we did not build pseudo-populations combining electrodes in the same labeled region across subjects. Instead, we used single electrodes and report average classification accuracy for single electrodes in Fig. 7B (whereas Fig. 4B is based on a pseudopopulation of 8 electrodes in

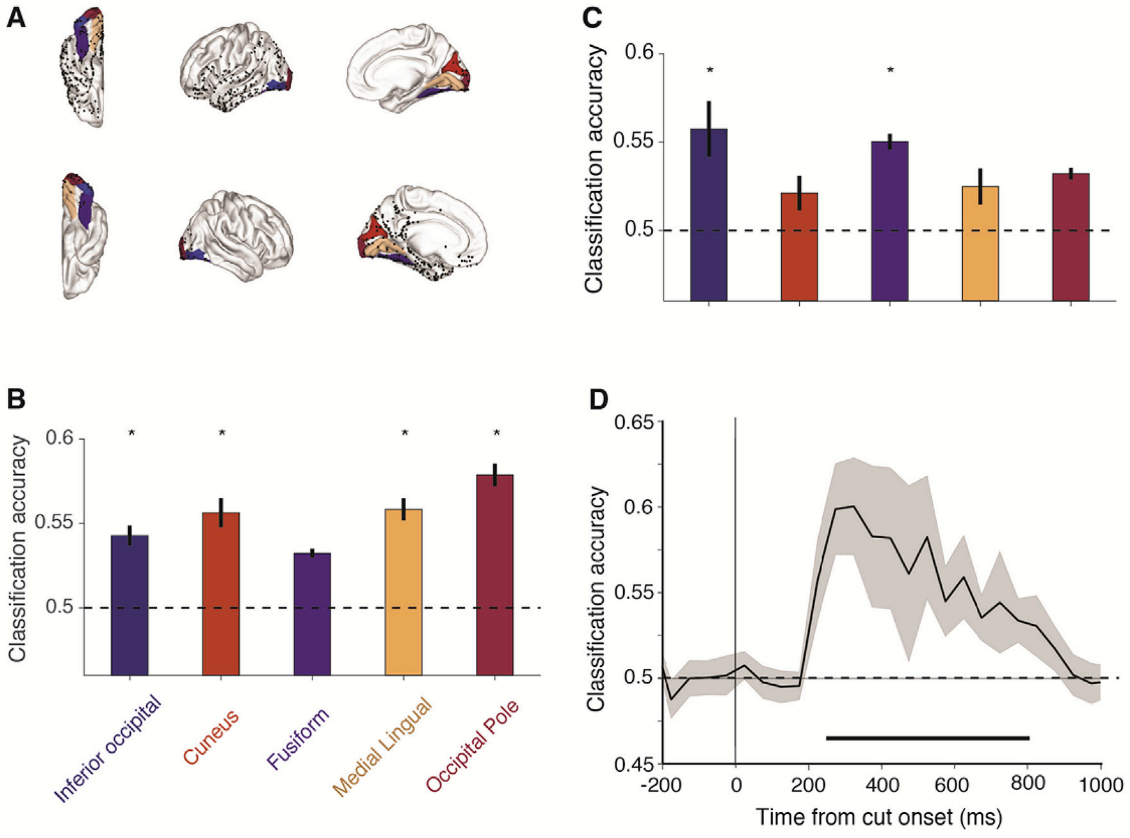


Fig. 7. Movie cuts can be decoded from a single presentation of a full-length movie. **A**. Location of all electrodes in Experiment II projected onto a common reference brain (Freesurfer fsaverage brain) shown at lateral and ventral views. Each dot corresponds to one electrode (total = 330 electrodes, Supplemental Table 1). The five anatomical regions (out of 20 regions with at least five electrodes) with significantly above chance classification accuracy in any of the decoding tasks in Fig. 7 are highlighted. **B**. Average single electrode classification accuracy between movie segments with a cut versus those without a cut for the 5 regions highlighted in Fig. 7A (mean ± SEM across all electrodes in each region). Chance = 0.5 (horizontal dashed line). The classification accuracy is calculated as the average from 50 to 400 ms post cut onset. The number of electrodes averaged is: inferior occipital, n = 7; cuneus, n = 10; fusiform, n = 26, medial lingual, n = 11, occipital pole, n = 13. Asterisks indicate regions with significantly above chance average classification accuracy based on a $p < 0.01$ permutation test (Methods). **C**. Average single electrode classification accuracy between movie shots with a face versus those without a face for those regions highlighted in Fig. 7A (mean ± SEM across all electrodes in each region). Chance = 0.5, horizontal dashed line. The classification accuracy is calculated as the average from 50 to 400 ms post cut onset. The number of electrodes averaged is the same as in Fig. 7B. Asterisks indicate regions with significantly above chance average classification accuracy based on a $p < 0.01$ permutation test. **D**. Visualization of dynamic classification accuracy for shots with a face versus those without a face versus time relative to cut onset using feature selection from all subjects and all electrodes (mean ± SEM across four subjects, Methods). Feature selection across all electrodes based on the training data only was applied at each time point to choose selective electrodes to be used in the classifier (Methods). Since the subjects viewed different movies, decoding results were then averaged post-hoc. Horizontal line indicates chance classification. The decoding was significantly above chance from 250 to 850 ms post-cut onset based on a $p < 0.01$ permutation test. See Fig. S11 for corresponding analyses in different frequency bands.

each region). Of the 20 regions with at least 5 electrodes, we observed a small but significant classification accuracy in 4 regions: inferior occipital gyrus, cuneus, medial lingual gyrus and occipital pole (see Figs. S11A–C for the corresponding analyses after filtering the IFP signals in different frequency bands).

Not all the same regions were interrogated in the different subjects that participated in Experiment I and II (Tables S4 and S5 provide detailed information about electrode locations in the two experiments). All of the 7 regions described in Fig. 4A had enough coverage to be considered in Experiment II. Three of these regions - inferior occipital, medial lingual gyrus and the occipital pole – showed significant classification accuracy to detect the presence of movie cuts in both experiments, while the other four regions did not reach significant classification accuracy in Experiment II. In addition, the cuneus showed significant classification accuracy to detect the presence of movie cuts in Experiment II but not in Experiment I.

3.7. *Invariant decoding of visual events in single presentation of movies (Experiment II)*

Following the steps in Experiment I, we next asked whether we could decode *what* changed in the image at a given movie cut. We trained the classifier to distinguish those shots containing a face from shots that did not contain a face following the procedures in Fig. 5, with two important differences. First, given the extensive preponderance of frames including human faces in the full-length movies in Experiment II, we labeled each shot as containing a face or not (as opposed to the animal faces in Experiment I, **Methods**). Second, as described above, we also considered single electrodes and report average classification accuracy in Fig. 7C, as opposed to results based on pseudopopulations. Of the 5 regions described in Fig. 7B, we could discriminate with small but significant classification accuracy shots containing a face from those with no face from single electrodes in the inferior occipital gyrus. Additionally, the fusiform gyrus also showed even smaller but still significant classification accuracy (see Figs. S11D–F for the corresponding analyses considering IFP signals filtered in different frequency bands).

To summarize the temporal dynamics in classification accuracy, we followed the procedure described in Figs. S8D and 5D for Experiment I and combined electrodes across all subjects in Fig. 7D. Again, because subjects watched different full-length movies, we did not combine electrodes across subjects but instead built pseudo-populations using each subjects' electrodes and averaged the four subjects' classification accuracies post-hoc. There was an increase in the classification accuracy to detect the presence or absence of a face starting slightly before 200 ms post cut onset and peaking around 300 ms post cut onset (Fig. 7D; see Figs. S11G–I for the corresponding analyses considering IFP signals filtered in different frequency bands). In sum, the previous section and this section demonstrate that the results obtained in Experiment I extrapolate to the conditions in Experiment II, whereby we can discriminate when there are visual changes and what those visual changes consist of in single presentations of a full-length movie.

4. Discussion

Parsing a continuous stream of visual stimuli is a fundamental challenge for the visual system. Here we considered commercial movies as a coarse proxy for natural visual input and described a methodology to extract visual information from invasive physiological recordings from the human brain during a continuous movie. Intracranial field potentials recorded along the ventral visual stream showed strong modulation approximately 100 ms after movie cuts, defined as discontinuous changes from one frame to the next (Fig. 2). Such vigorous physiological responses allowed us to detect *when* there are visual changes during the continuous stimulus (Fig. 4B–D). By aligning the responses to those changes, we identified *what* visual information was present in each shot (e.g., shots with or without an animal), generalizing across different events within the same movie or even across different movies (Fig. 5). We further demonstrated that these findings extend to detecting the timing of visual changes and decoding events in a single presentation of a full-length movie (Figs. 6–7).

We separately considered broadband signals from 0.1 to 100 Hz and broadband, band-limited signals in the alpha (8–15 Hz), low gamma (25–70 Hz) and high gamma (70–120 Hz) bands. We observed fewer and weaker visual responses in the alpha band, consistent with previous studies (e.g. Bansal et al., 2012). The qualitative and conceptual conclusions derived from examining the low and high gamma band were consistent with those based on the broadband signals. Yet, there were quantitative differences in terms of the numbers of responsive electrodes, classification performance and, in some cases, the specific areas that showed significant decoding accuracy. These differences are discussed in further detail in the Supplementary Material. These qualitative similarities and quantitative differences between broadband and gamma band responses were noted in several previous studies (e.g. (Vidal et al., 2010; Privman et al., 2011; Bansal et al., 2012; Miller et al., 2014)).

Commercial movies such as the ones used here and in other studies clearly constitute artificial stimuli that are different from natural viewing conditions. Movies are commercial forms of art specifically and carefully designed to evoke strong emotional experiences, producing memorable audiovisual scenes in a compressed time frame beyond the occurrences of everyday life. Movie cuts are introduced in videos by the director to manipulate spatial coordinates, context, attention, and interactions (Dudai, 2012; Smith et al., 2012). These cuts only constitute a first order approximation to the type of discontinuities that arise under natural viewing conditions as a result of sudden changes in moving objects, occlusion, lighting and internally dictated changes such as eye movements. Despite these caveats, movies provide a rich stimulus for probing neural responses in situations where the brain is continuously subject to incoming inputs, as opposed to a blank screen followed by the onset of a picture. Indeed, several previous studies have demonstrated that sharp transitions between frames in movies can trigger a strong neural response all along ventral visual cortex from early visual areas (Vinje and Gallant, 2000; Montemurro et al., 2008) to the highest visual areas (Privman et al., 2007; Honey et al., 2012; McMahon et al., 2015).

Critically, the brain must be able to capture these dynamic transitions in single events without averaging responses over multiple repetitions. Even with the type of coarse signals and limited spatial sampling considered here, it is possible to detect visual changes in a movie within approximately 100 ms of those changes (Figs. 2, 4 and 6). These latencies are close to those reported in monkey and human ventral visual cortex in response to static images (Richmond et al., 1990; Rolls and Tovee, 1995; Keysers et al., 2001; Hung et al., 2005; Liu et al., 2009). Thus, our intuitions about the initial dynamics of neural responses triggered by flashing static pictures seem to extrapolate to dynamic and continuous viewing conditions.

The rapid field potential changes were elicited by most movie cuts and were consistent throughout tens of repetitions. Intriguingly, we observed few consistent physiological responses across repetitions outside of movie cuts (Fig. 2D and Fig. 3). In other words, we largely failed to note consistent responses from one repetition of the movie clip to another except within a few hundred milliseconds after a movie cut. There are several non-exclusive possibilities for this observation. First, our sampling of brain locations was far from exhaustive. The electrode locations were strictly dictated by clinical criteria. Although we interrogated a relatively large number of brain regions for this type of study (almost 1000 different electrodes distributed over 46 brain regions, Table S4), there could well be many other brain loci that show consistent responses to other aspects of the movies unrelated to the movie cuts. Second, we studied coarse field potential signals recorded from low-impedance electrodes that capture neural activity over vast numbers of neurons (Buzsáki et al., 2012). It is quite possible that there are strong neuronal responses to other aspects of the movies that are not captured by field potential signals. Third, it is conceivable that other aspects of cognition beyond visual processing are modulated or even governed by different mechanisms that do not lead to the type of sharp and consistent responses illustrated in Fig. 2. In particular, other aspects of cognition beyond visual processing during a movie may not have a well-defined temporal onset (e.g. when exactly emotions are triggered during a scene), or they may show rapid adaptation (e.g. the first viewing of a movie scene might trigger stronger emotions than the tenth viewing), both of which would reduce the reproducibility of these signals across multiple trials. In sum, while we argue here that we can rapidly decode visual transitions in single events during a movie, there remain important questions about how to study the neural basis of higher cognitive functions under natural conditions.

In the absence of fixed image onset times or movie cuts, the brain must segment continuous information into discrete visual events. How are visually evoked signals aligned under natural viewing conditions? Several sources in the brain could in principle provide an internal alignment signal to the ventral visual stream, including a copy of a motor efferent from eye movements, or external object movement onset information conveyed by the dorsal stream. While this study does not explain the mechanistic origin for the physiological changes triggered by movie cuts, the results presented here show that it is possible to align and interpret signals directly from the field potentials recorded from electrodes in the ventral visual stream. During natural viewing conditions, we speculate that signals along the ventral visual stream may be sufficient to interpret what changes when without the need for additional sources of information.

The main regions along the ventral visual stream that contributed to decoding when and what information included the inferior occipital gyrus, the fusiform gyrus, the inferior temporal gyrus and the occipital pole (Figs. 4 and 7). All of these regions have also revealed selective visual responses in previous invasive human neurophysiology studies (e.g. (Privman et al., 2007; Liu et al., 2009; Vidal et al., 2010)). These areas are also consistent with locations highlighted in non-invasive human fMRI studies (e.g (Grill-Spector and Malach, 2004).,) and with putative homologous regions in the macaque brain (e.g. (Logothetis and Sheinberg, 1996; Tanaka, 1996; Connor et al., 2007)).

Once the onset of visual changes is detected, approximately the same ventral visual regions provide a rich representation that contains selective information about the nature of those changes (Figs. 5 and 7C-D). Selective visual information arose within the first 200 ms of a movie cut, and was relatively robust to the many highly varied transformations that took place in these commercial movies. Specifically, in Experiment I, classifiers trained to detect the presence versus absence of a humanized animal, using electrodes in the inferior occipital gyrus, inferior temporal gyrus or fusiform gyrus, showed a significant degree of extrapolation to independent test data from a completely different movie clip (Fig. 5C, green bars). In Experiment II, classifiers trained to discriminate the presence versus absence of human faces from the field potential responses from single electrodes in the inferior occipital gyrus or fusiform gyrus showed a weak but significant degree of extrapolation to independent test data during single repetitions of other parts of the movie (Fig. 7C–D). The results in Fig. 5 should *not* be interpreted to imply that those electrodes were selective to “humanized animals” or that the corresponding analyses in Fig. 7 imply selectivity for “human faces”. This study used commercial movies and no attempt was made to circumscribe the visual changes to the appearance of animals or faces. The appearance of animals and faces was correlated and accompanied by changes in motion, contrast and many other visual properties. It seems likely that the main drivers of the strong visually evoked transitions, such as the ones illustrated in Fig. 2, are the sharp contrast changes and motion energy changes triggered by movie cuts. Further studies directly comparing the responses to dynamic stimuli versus stimulus flashes will be needed to further dissect the specific features that dictate selectivity to movie events revealed here. The current results demonstrate that it is possible to distill reliable, selective and invariant information, even in single events during a continuous stream of frames.

Moving from repeated presentations of identical, static stimuli with fixed onsets and offsets to movie stimuli constitutes an important step to bridge the gap between laboratory studies and understanding vision in the real world. Furthermore, movies present rich visual and social input. The initial methodological steps suggested here open the doors to interpreting neural responses to complex cognitive events during single presentations of movies.

Acknowledgements

We thank all of the patients for participating in this study. This paper is based upon work supported by NIH (R01EY026025) and the Center for Brains Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.neuroimage.2017.08.027.

References

Bansal, A.K., Singer, J.M., Anderson, W.S., Golby, A., Madsen, J.R., Kreiman, G., 2012. Temporal stability of visually selective responses in intracranial field potentials recorded from human occipital and temporal lobes. J. Neurophysiol. 108, 3073–3086.

Bartels, A., Zeki, S., 2005. Brain dynamics during natural viewing conditions—a new guide for mapping connectivity in vivo. Neuroimage 24, 339–349.

Bokil, H., Andrews, P., Kulkarni, J.E., Mehta, S., Mitra, P.P., 2010 Sep 30. Chronux: a platform for analyzing neural signals. J. Neurosci. Methods 192 (1), 146–151.

Booth, M.C., Rolls, E.T., 1998. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex 8, 510–523.

Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. 13, 407–420.

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., Rust, N.C., 2005. Do we know what the early visual system does? J. Neurosci. 25.

Connor, C.E., Brincat, S.L., Pasupathy, A., 2007. Transformation of shape information in the ventral pathway. Curr. Opin. Neurobiol. 17, 140–147.

Conroy, B.R., Singer, B.D., Guntupalli, J.S., Ramadge, P.J., Haxby, J.V., 2013. Intersubject alignment of human cortical anatomy using functional connectivity. Neuroimage 81, 400–411.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15.

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434.

Dudai, Y., 2012. The cinema-cognition dialogue: a match made in brain. Front. .Hum. Neurosci. 6, 248.

Felsen, G., Dan, Y., 2005. A natural approach to studying vision. Nat. Neurosci. 8, 1643–1646.

Fiser, J., Chiu, C., Weliky, M., 2004. Small modulation of ongoing cortical dynamics by sensory input during natural vision. Nature 431, 573–578.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640.

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, 423–434.

Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J., 2005. Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863–866.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.

Isik, L., Meyers, E.M., Leibo, J.Z., Poggio, T., 2014. The dynamics of invariant object recognition in the human visual system. J. Neurophysiol. 111, 91–102.

Keysers, C., Xiao, D.-K., Földiák, P., Perrett, D.I., 2001. The speed of sight. J. Cognit. Neurosci. 13, 90–101.

Kriegeskorte, N., Kreiman, G., 2011. Visual Population Codes: toward a Common Multivariate Framework for Cell. Google Books.

Lei, Y., Sun, N., Wilson, F.A.W., Wang, X., Chen, N., Yang, J., Peng, Y., Wang, J., Tian, S., Wang, M., Miao, Y., Zhu, W., Qi, H., Ma, Y., 2004. Telemetric recordings of single neuron activity and visual scenes in monkeys walking in an open field. J. Neurosci. Methods 135, 35–41.

Lewen, G.D., Bialek, W., Steveninck, RR d. R v, 2001. Neural coding of naturalistic motion stimuli. Netw. Comput. Neural Syst. 12, 317–329.

Liu, H., Agam, Y., Madsen, J.R., Kreiman, G., 2009. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. Neuron 62, 281–290.

Logothetis, N.K., Sheinberg, D.L., 1996. Visual object recognition. Annu. Rev. Neurosci. 19, 577–621.

McMahon, D.B.T., Russ, B.E., Elnaiem, H.D., Kurnikova, A.I., Leopold, D.A., 2015. Single-unit activity during natural vision: diversity, consistency, and spatial sensitivity among AF face patch neurons. J. Neurosci. 35.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., Poggio, T., 2008. Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol. 100, 1407–1419.

Miller, K.J., Honey, C.J., Hermes, D., Rao, R.P., denNijs, M., Ojemann, J.G., 2014. Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. Neuroimage 85, 711–720.

Montemurro, M.A., Rasch, M.J., Murayama, Y., Logothetis, N.K., Panzeri, S., 2008. Phase-of-firing coding of natural visual stimuli in primary visual cortex. Curr. Biol. 18 (5), 375–380.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646.

Ojemann, G.A., 1997. Treatment of temporal lobe epilepsy. Annu. Rev. Med. 48, 317–328.

Podvalny, E., Yeagle, E., Mégevand, P., Sarid, N., Harel, M., Chechik, G., Mehta, A.D., Malach, R., 2016. Invariant temporal dynamics underlie perceptual stability in human visual cortex. Curr. Biol. 27 (2), 155–165.

Privman, E., Fisch, L., Neufeld, M.Y., Kramer, U., Kipervasser, S., Andelman, F., Yeshurun, Y., Fried, I., Malach, R., 2011. Antagonistic relationship between gamma power and visual evoked potentials revealed in human visual cortex. Cerebr. Cortex 21, 616–624.

Privman, E., Nir, Y., Kramer, U., Kipervasser, S., Andelman, F., Neufeld, M.Y., Mukamel, R., Yeshurun, Y., Fried, I., Malach, R., 2007. Enhanced category tuning revealed by intracranial electroencephalograms in high-order human visual areas. J. Neurosci. 27.

Richmond, B.J., Optican, L.M., Spitzer, H., 1990. Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex. I. Stimulus-response relations. J. Neurophysiol. 64.

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025.

Rolls, E.T., Tovee, M.J., 1995. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. 73.

Russ, B.E., Leopold, D.A., 2015. Functional MRI mapping of dynamic visual features during natural viewing in the macaque. Neuroimage 109, 84–94.

Rust, N.C., Movshon, J.A., 2005. In praise of artifice. Nat. Neurosci. 8, 1647–1650.

Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. U. S. A. 104, 6424–6429.

Smith, T.J., Levin, D., Cutting, J.E., 2012. A window on reality. Curr. Dir. Psychol. Sci. 21, 107–113.

Tanaka, K., 1996. Inferotemporal cortex and object vision. Annu. Rev. Neurosci. 19, 109–139.

Tang, H., Buia, C., Madhavan, R., Crone, N.E., Madsen, J.R., Anderson, W.S., Kreiman, G., 2014. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. Neuron 83, 736–748.

Vidal, J.R., Ossandón, T., Jerbi, K., Dalal, S.S., Minotti, L., Ryvlin, P., Kahane, P., Lachaux, J.-P., 2010. Category-specific visual responses: an intracranial study comparing gamma, beta, alpha, and ERP response selectivity. Front. Hum. Neurosci. 4, 195.

Vinje, W.E., Gallant, J.L., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287 (80-).

Whittingstall, K., Bartels, A., Singh, V., Kwon, S., Logothetis, N.K., 2010. Integration of EEG source imaging and fMRI during continuous viewing of natural movies. Magn. Reson. Imaging 28, 1135–1142.