Journal of Neural Engineering

PURPOSE-LED PUBLISHING™

PAPER

Identifying temporal correlations between natural single-shot videos and EEG signals

To cite this article: Yuanyuan Yao *et al* 2024 *J. Neural Eng.* **21** 016018

View the article online for updates and enhancements.

Journal of Neural Engineering

RECEIVED
15 September 2023

REVISED
15 January 2024

ACCEPTED FOR PUBLICATION
26 January 2024

PUBLISHED
7 February 2024

PAPER

Identifying temporal correlations between natural single-shot videos and EEG signals

Yuanyuan Yao[1], Axel Stebner[2], Tinne Tuytelaars[2], Simon Geirnaert[3] and Alexander Bertrand[1,*]

[1] Department of Electrical Engineering, STADIUS, KU Leuven, Leuven, Belgium
[2] Department of Electrical Engineering, PSI, KU Leuven, Leuven, Belgium
[3] Department of Electrical Engineering, STADIUS, Department of Neurosciences, ExpORL, KU Leuven, Leuven, Belgium
[*] Author to whom any correspondence should be addressed.

E-mail: alexander.bertrand@esat.kuleuven.be

Keywords: BCI, EEG decoding, video stimuli

Abstract

Objective. Electroencephalography (EEG) is a widely used technology for recording brain activity in brain-computer interface (BCI) research, where understanding the encoding-decoding relationship between stimuli and neural responses is a fundamental challenge. Recently, there is a growing interest in encoding-decoding natural stimuli in a single-trial setting, as opposed to traditional BCI literature where multi-trial presentations of synthetic stimuli are commonplace. While EEG responses to natural speech have been extensively studied, such stimulus-following EEG responses to natural video footage remain underexplored. *Approach.* We collect a new EEG dataset with subjects passively viewing a film clip and extract a few video features that have been found to be temporally correlated with EEG signals. However, our analysis reveals that these correlations are mainly driven by shot cuts in the video. To avoid the confounds related to shot cuts, we construct another EEG dataset with natural single-shot videos as stimuli and propose a new set of object-based features. *Main results.* We demonstrate that previous video features lack robustness in capturing the coupling with EEG signals in the absence of shot cuts, and that the proposed object-based features exhibit significantly higher correlations. Furthermore, we show that the correlations obtained with these proposed features are not dominantly driven by eye movements. Additionally, we quantitatively verify the superiority of the proposed features in a match-mismatch task. Finally, we evaluate to what extent these proposed features explain the variance in coherent stimulus responses across subjects. *Significance.* This work provides valuable insights into feature design for video-EEG analysis and paves the way for applications such as visual attention decoding.

1. Introduction

Electroencephalography (EEG) is a non-invasive technology to record the electrical activity of the brain through electrodes attached to the scalp. Due to its high temporal resolution, affordability, and portability, EEG has found extensive applications in brain-computer interface (BCI) research. A key challenge in BCIs is understanding the encoding-decoding relationship between stimuli and EEG responses. Early approaches, e.g. the event-related potential (ERP) approach [1], heavily relied on short synthetic sensory stimuli, such as tone beeps or sudden visual events. Participants were repeatedly presented with the same synthetic stimulus, and the neural response was obtained by averaging the EEG signals across trials to deal with the low signal-to-noise ratio (SNR) of EEG. Using synthetic stimuli leads to more deterministic neural responses, such that averaging multiple EEG trials allows to remove the uncorrelated background noise while retaining the neural signals of interest. However, such multi-trial paradigms often lead to participant fatigue and are impractical for real-life applications involving natural continuous stimuli like audio and video footage that emerge in everyday life applications, in which case the stimuli are only presented once. Such natural settings are often assumed in passive BCIs where the goal is, e.g. to monitor attention or engagement [2, 3], to detect mental fatigue and workload [4], or to recognize emotions such as stress, surprise, and fear [5]. Consequently, there has been a growing need to develop new paradigms that allow to decode neural responses to a natural stimulus in a single-trial context. The challenge lies in the fact that the stimulus-following responses are non-deterministic (as opposed to ERPs), and that they are buried under all kinds of EEG background noise, requiring more advanced signal processing tools to decode them, in particular since multi-trial averaging is not an option. The latter implies that often a longer trial duration is required in a single-trial setting, in order to obtain sufficiently reliable decoding results.

Therefore, when working with natural single-trial stimuli, two crucial elements are (1) a good (feature) representation of the stimulus that correlates well with the EEG and (2) an appropriate model that captures the relationship between the stimulus representation and the EEG response, while removing background EEG. Extensive research has been conducted on auditory-EEG analysis for natural speech, where various useful speech representations have been proposed, ranging from low-level features such as the spectrogram [6] and speech envelope [7–9], to high-level information such as phonemes [10] and semantic context [11]. The most common models are linear models, which can be roughly categorized into two groups: forward models and backward models. Linear forward models assume that the EEG signal consists of a stimulus response superimposed to background EEG, where the former is typically modelled as a convolution between a proper stimulus representation (e.g. a speech envelope) and a so-called temporal response function (TRF). The TRF can be estimated using, e.g. least squares (LS), and the EEG signals can be predicted from the audio features using the estimated TRF [7, 10, 11]. Backward models, on the other hand, reconstruct the stimulus as a linear combination of (lagged) EEG channels [8, 9]. A hybrid encoding-decoding model based on canonical correlation analysis (CCA) was proposed in [12], where linear transformations were applied to both the speech envelope and the EEG signals such that the latent representations were maximally correlated. However, linear models have an inherent limitation in capturing the nonlinear dynamics of the brain. Moreover, using linear models also makes the results very dependent on the handcrafted feature engineering of the stimulus representation. Therefore, deep learning methods have been receiving increasing attention in recent years [13]. For instance, [14] decoded the auditory brain using deep learning-based CCA, and in [15], a long short-term memory-based model was proposed to discriminate whether a pair of an EEG segment and speech envelope correspond to each other or not. One direct application of audio-EEG analysis is auditory attention decoding, which opens the doors for advancing future technologies such as neuro-steered hearing aids [16].

While there have been successful attempts to decode natural audio from EEG, the decoding of natural video footage from EEG has received less attention. The high-dimensional nature of the video signals poses challenges in finding useful representations. A possible approach is to not explicitly take the video stimulus into account in the modeling and extract common EEG components across the EEGs of multiple subjects watching the same video using methods such as correlated component analysis (CorrCA) [17–20]. By construction, the EEG responses that are coherent across subjects can only be time-locked to the visual stimuli since only the video stimuli is shared during all the EEG measurements. However, the link between the extracted EEG components and the video is unclear, making these components not very interpretable, in particular in regard to which features in the video drive the correlation. Alternatively, stimulus-aware algorithms such as CCA can be used to analyze the encoding-decoding relationship between the (specific features of) video stimulus and individual EEG signals. In this case, the design of relevant (low-dimensional) video features becomes crucial. In previous studies, the mean velocity of pixels calculated from the optical flow and the mean temporal derivative of pixel intensity (temporal contrast) were shown to be correlated with individual EEG signals, suggesting that they could be visual features that elicit strong EEG responses [21, 22].

In this work, we argue that shot cuts, which refer to sudden changes in the camera viewpoint or scene, have a significant impact on stimulus-aware video-EEG analysis, using previously proposed video features, as well as on (stimulus-unaware) multi-subject EEG analysis. Moreover, we recorded a new EEG dataset with subjects watching a set of single-shot videos containing a single moving object (i.e. a person). These single-shot single-object videos were specifically chosen to avoid introducing confounds related to shot cuts and to reduce the complexity of the stimuli. We demonstrate that the previously proposed mean velocity of pixels and temporal contrast are not relevant enough to generate significant correlations in the absence of shot cuts, and propose new object-based versions of these features that are more relevant, leading to significantly higher correlations. Moreover, we show that the EEG components obtained with the new features are not dominantly driven by eye movements, which are usually considered confounds. We further demonstrate that the proposed features are of better quality by their lower error rates in a match-mismatch (MM) task. Finally, we perform a multi-subject EEG analysis and calculate the proportion of variance in the coherent stimulus responses explained by the proposed features.

The structure of the paper is outlined as follows: section 2 describes the experimental protocol in detail. In section 3, we introduce the proposed video features and review the mathematical tools that we used in our analysis. The results are presented in section 4, followed by further discussions in section 5. Finally, in section 6, we draw conclusions based on our findings.

2. Experiment

2.1. Subjects and stimuli

20 young, healthy participants were recruited (10 females) for this study. 14 single-shot videos (duration 202 s–463 s) showing a single moving person during various activities (dancing, mime, acrobatics, magic shows) were selected from YouTube, which were then concatenated into two longer trials (duration 36 min and 35 min) in a sequence that ensured diverse content, e.g. a dance performance followed by a mime show and a magic show, to reduce participants' fatigue. Smooth transitions with cross-fading effects were applied between the videos. 10 of the subjects (5 females) watched an extra 24 min clip from Mr Bean, which contains shot cuts distributed irregularly throughout the video, and which was used to study the effect of shot cuts. For clarity, we will refer the first dataset as the Single-Shot dataset, and the second as the MrBean dataset. Unless mentioned otherwise, the same processing steps are applied to both datasets. A squared box that flashed once every 30 s was encoded into the videos for synchronization. It appeared outside of the original frames and was positioned in the top right corner of the screen. During passive viewing, the flashing box was covered such that it was not distractive. To avoid the confounds introduced by audio, all the videos were muted during the experiment. The study was approved by the KU Leuven Social and Societal Ethics Committee, and before participating, all participants provided their informed consent by signing a consent form.

2.2. Data acquisition and preprocessing

The EEG data of these subjects was recorded when they were watching these videos. These experiments were conducted in a quiet and dark room to minimize potential distractions. The subjects were instructed to watch the videos naturally, concentrate on the content, and minimize their own movements. The EEG data was recorded with a biosemi activetwo system at a sample rate of 2048 Hz. The participants wore a 64-channel EEG cap, and 4 electrooculogram (EOG) sensors around the eyes were used to track eye movements. Potential bad channels were carefully logged during each session. A photo sensor for detecting the light changes of the embedded flashing box was also connected to the recorder to provide synchronization information. The collected EEG data was first segmented into different pieces corresponding to each video based on the signal of the photo sensor.

The video preprocessing involved resampling to 30 Hz and resizing to 854×480 pixels. Features were extracted frame by frame as detailed in section 3.1. Preprocessing of the EEG data was performed using functions from the MNE-Python library [23], which involved the following steps: (1) interpolating bad channels; (2) re-referencing the data to the average of all channels; (3) applying a high pass filter with a cutoff frequency of 0.5 Hz; (4) removing power line noise with a notch filter; (5) downsampling the data to 30 Hz (to match the video frame rate) after proper anti-aliasing filtering; and (6) regressing out the EOG channels to reduce eye artifacts. The specifications of the filters can be found in appendix A. Importantly, the filters are zero phase and thus do not introduce delays.

To avoid potential confounding effects caused by the (cross-fading) transitions between consecutive single-shot videos, the initial and final second of each video, along with the corresponding EEG data, were excluded from the analysis. Additionally, due to synchronization issues for one subject in the Single-Shot dataset, one of the single-shot videos was excluded for all subjects, resulting in a dataset of 20 subjects with 63 min of data each. Similarly, for the MrBean dataset, the initial and final second of the data were discarded to avoid any influence from video onset and termination. The MrBean dataset included 10 subjects, each contributing 24 min of data.

3. Methods

In section 3.1, we explain the newly proposed video feature that we use in the video-EEG analysis. To quantify and identify the temporal coupling between signals, we choose CCA in the context of stimulus-aware video-EEG analysis, to find the correlation between the individual EEG signals and the video features (section 3.2), while using multi-set extensions of CCA such as generalized canonical correlation analysis (GCCA) and CorrCA in the context of multi-subject EEG analysis, to find the correlation between EEG signals of multiple subjects watching the same video (section 3.3). We briefly review these methods and point out the links between their seemingly-different original formulations. Sections 3.4–3.6 delve into the details of hyperparameters used in these algorithms, evaluation metrics, and the interpretation of obtained filter weights, respectively. The code for extracting features, implementing algorithms, and reproducing results can be found at: https://github.com/YYao-42/Identifying-Temporal-Correlations-Between-Natural-Single-shot-Videos-and-EEG-Signals.

In the subsequent discussion, we consider the C-channel EEG signals $\mathbf{x}_n(t) \in \mathbb{R}^C$ recorded from N

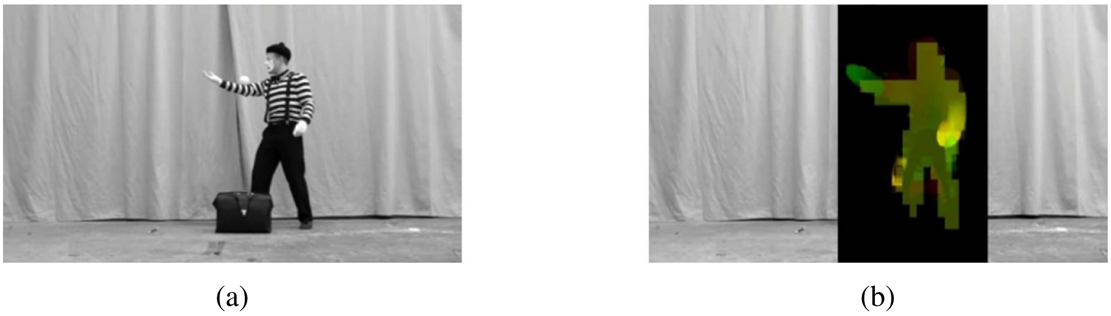


Figure 1. (a) A frame of a video that was used in our experiment. (b) We can obtain the bounding box and the segmentation masks using Mask R-CNN. The area within the bounding box is in black. We overlay the segmentation masks on top of it, and the color is related to the value of features.

subjects watching the same video stimulus, where t is the time index, and n is an index that refers to a specific subject ($n \in 1,\ldots,N$). The stimulus is represented by $\mathbf{y}(t) \in \mathbb{R}^{D_y}$, a D_y-dimensional time-dependent feature extracted from the video, where the time index t is the same as the time index in the EEG signals. Without loss of generality, we assume that $\mathbf{x}_n(t)$ and $\mathbf{y}(t)$ are zero-mean, i.e. $\mathbb{E}\{\mathbf{x}_n(t)\} = \mathbf{0}$ and $\mathbb{E}\{\mathbf{y}(t)\} = \mathbf{0}$. Furthermore, we assume the availability of T time samples, leading to the data matrices $\mathbf{X}_n \in \mathbb{R}^{T\times C}$ and $\mathbf{Y} \in \mathbb{R}^{T\times D_y}$, where each row is a sample of the EEG signal and the feature, respectively.

3.1. Video feature extraction

3.1.1. Optical flow and temporal contrast

In the limited literature on video-EEG analysis, it has been observed that optical flow and temporal contrast can be correlated with EEG signals [21]. Optical flow estimates the velocity vectors of pixels between consecutive frames and can thus be used to capture the motion information in videos. We applied the Gunnar-Farneback Optical Flow algorithm implemented in OpenCV to extract the flow vectors [24, 25], and computed the magnitude of each velocity vector $|\mathbf{v}_m(\mathbf{z})|$, where the subscript m denotes the frame index and $\mathbf{z}$ denotes the pixel coordinate. Temporal contrast is a low-level feature that is simply defined as the absolute intensity changes between consecutive frames, i.e. $\Delta I_m(\mathbf{z}) = |I_m(\mathbf{z}) - I_{m-1}(\mathbf{z})|$. In [21], both types of features are averaged across all pixels, resulting in a scalar value for each frame. We refer to the average magnitude of velocity as *AvgFlow* and the average intensity change as *AvgTempCtr*.

3.1.2. Newly proposed object-based features

In the previous approach of averaging over all pixels, important spatial information is lost, making it impossible to identify specific regions in a frame that elicit strong neural responses. Additionally, treating all pixels equally may not accurately reflect the attentional focus of the subject, as certain pixels, such as those corresponding to moving objects, are more likely to attract attention compared to the background, while the latter can still have a large impact on optical flow or temporal contrast. Moreover, the scaling of objects also influences the results. For example, consider an object moving at a constant speed but changing in scale. In this case, its contribution to the result varies as the number of pixels changes, which is not ideal for capturing the neural responses related to movement perception. To overcome these limitations, we propose a refined version that incorporates object segmentation. This approach involves segmenting the object(s) in each frame and calculating the mean values only for the pixels belonging to the object(s).

To perform object segmentation, we utilize Mask R-CNN [26], a deep learning model designed for object detection and segmentation in images. This model consists of two branches: one branch returns the bounding boxes of the detected objects, while the other branch provides segmentation masks. The segmentation masks are matrices where each entry indicates whether a pixel belongs to an object (1) or not (0). By feeding frames directly into the pre-trained Mask R-CNN model, we obtain the segmentation masks, which allow us to identify the pixels associated with each object. An example is shown in figure 1.

With the obtained segmentation masks, we selectively average the flow vectors and unsigned intensity changes only over the pixels belonging to the identified object(s). In the multi-object case, one could average over the union of all pixels across all objects, or alternatively, normalize for each object separately and then sum the per-object features. A pre-selection of relevant objects could also be performed before feature fusion based on, e.g. the sizes of the bounding boxes. To circumvent feature fusion, one could treat objects separately in the analysis. Due to the many additional degrees of freedom, the multi-object case is beyond the scope of this paper and we focus on the single-object case. This object-based version of optical flow and temporal contrast is referred to as *ObjFlow* and *ObjTempCtr*, respectively. A summary of the feature definitions is given in table 1. By incorporating object segmentation, we can retain the spatial information and stress more the regions that may receive more attention and presumably elicit higher

Table 1. Video features and definitions. $\mathcal{Z}$ and $\mathcal{O}$ denote the set of all pixels and the set of pixels belonging to the object of interest, respectively. $|\mathcal{Z}|$ and $|\mathcal{O}|$ represent the cardinality of the sets. $|\mathbf{v}_m(\mathbf{z})|$ is the magnitude of velocity.

Feature	Abbreviation	Definition
Average optical flow [21]	*AvgFlow*	$\frac{1}{\lvert\mathcal{Z}\rvert}\sum_{z\in\mathcal{Z}}\lvert\mathbf{v}_m(\mathbf{z})\rvert$
Average temporal contrast [21]	*AvgTempCtr*	$\frac{1}{\lvert\mathcal{Z}\rvert}\sum_{z\in\mathcal{Z}}\Delta I_m(\mathbf{z})$
Object-based optical flow	*ObjFlow*	$\frac{1}{\lvert\mathcal{O}\rvert}\sum_{z\in\mathcal{O}}\lvert\mathbf{v}_m(\mathbf{z})\rvert$
Object-based temporal contrast	*ObjTempCtr*	$\frac{1}{\lvert\mathcal{O}\rvert}\sum_{z\in\mathcal{O}}\Delta I_m(\mathbf{z})$

neural responses, providing a more refined feature representation. Note that this method compensates for the variation in the number of pixels within the object since the values are averaged over pixels of the identified object.

3.2. Stimulus-aware video-EEG analysis: CCA

For conciseness, we drop the subscript n of the EEG signals in this subsection, as the following analysis is made per subject individually. CCA was proposed in [27] as a method to find relations between two sets of variables. Given multi-dimensional EEG signal $\mathbf{x}(t)$ and video feature $\mathbf{y}(t)$, CCA computes filters $\mathbf{w}_x \in \mathbb{R}^C$ (on the EEG) and $\mathbf{w}_y \in \mathbb{R}^{D_y}$ (on the stimulus) that maximize the Pearson correlation coefficient between the filtered output signals (figure 2(a)). Therefore, CCA can be formulated as the following optimization problem:

$$\underset{\mathbf{w}_x,\mathbf{w}_y}{\text{maximize}} \quad \frac{\mathbb{E}\left\{\left[\mathbf{w}^{\mathrm{T}}{}_x\mathbf{x}(t)\right]\left[\mathbf{w}^{\mathrm{T}}{}_y\mathbf{y}(t)\right]\right\}}{\sqrt{\mathbb{E}\left\{\left[\mathbf{w}^{\mathrm{T}}{}_x\mathbf{x}(t)\right]^2\right\}}\sqrt{\mathbb{E}\left\{\left[\mathbf{w}^{\mathrm{T}}{}_y\mathbf{y}(t)\right]^2\right\}}}. \tag{1}$$

The optimal $\mathbf{w}_x$ and $\mathbf{w}_y$ are called the first canonical components, and the transformed signals $\mathbf{w}^{\mathrm{T}}{}_x\mathbf{x}(t)$ and $\mathbf{w}^{\mathrm{T}}{}_y\mathbf{y}(t)$ are called the first canonical directions. Here we assume the filters only act on the current timestamp, but the formulation can be easily generalized to incorporate temporal information. This can be accomplished by extending $\mathbf{x}(t)$ with $L_x - 1$ time-lagged copies of $\mathbf{x}(t)$, such that it becomes a CL_x-dimensional vector, and extending $\mathbf{y}(t)$ with $L_y - 1$ time lagged copies of $\mathbf{y}(t)$, resulting in a D_yL_y-dimensional vector (similar for $\mathbf{w}_x$ and $\mathbf{w}_y$). Such extension can also compensate for the unknown time lag between the stimulus and the EEG response.

Since the scaling of $\mathbf{w}_x$ and $\mathbf{w}_y$ does not affect the objective function in (1), we can constrain the canonical directions to have unit variance to simplify the denominator. By denoting correlation matrices $\mathbf{R}_{xy} = \mathbb{E}\{\mathbf{x}(t)\mathbf{y}(t)^{\mathrm{T}}\} \in \mathbb{R}^{C\times D_y}$, $\mathbf{R}_{xx} = \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}\} \in \mathbb{R}^{C\times C}$, and $\mathbf{R}_{yy} = \mathbb{E}\{\mathbf{y}(t)\mathbf{y}(t)^{\mathrm{T}}\} \in \mathbb{R}^{D_y\times D_y}$ (which can be estimated as $\frac{1}{T}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, $\frac{1}{T}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\frac{1}{T}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, respectively), (1) can be rewritten as a constrained optimization problem:

$$\begin{aligned} \underset{\mathbf{w}_x,\mathbf{w}_y}{\text{maximize}} \quad & \mathbf{w}^{\mathrm{T}}{}_x\mathbf{R}_{xy}\mathbf{w}_y \\ \text{subject to} \quad & \mathbf{w}^{\mathrm{T}}{}_x\mathbf{R}_{xx}\mathbf{w}_x = 1, \\ & \mathbf{w}^{\mathrm{T}}{}_y\mathbf{R}_{yy}\mathbf{w}_y = 1. \end{aligned} \tag{2}$$

An extension of (2) in the multi-component case is:

$$\begin{aligned} \underset{\mathbf{W}_x,\mathbf{W}_y}{\text{maximize}} \quad & \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}{}_x\mathbf{R}_{xy}\mathbf{W}_y\right) \\ \text{subject to} \quad & \mathbf{W}^{\mathrm{T}}{}_x\mathbf{R}_{xx}\mathbf{W}_x = \mathbf{I}_K, \\ & \mathbf{W}^{\mathrm{T}}{}_y\mathbf{R}_{yy}\mathbf{W}_y = \mathbf{I}_K, \end{aligned} \tag{3}$$

where $\mathrm{Tr}(\cdot)$ is the trace operator, K is the number of components, $\mathbf{I}_K$ is the $K \times K$ identity matrix, and the k-th columns of $\mathbf{W}_x \in \mathbb{R}^{C\times K}$ and $\mathbf{W}_y \in \mathbb{R}^{D_y\times K}$ are the k-th canonical components. The constraints require that the canonical directions of different orders are uncorrelated to avoid trivial solutions. It can be shown that (3) can be written in a more compact form (up to a scaling factor in the solution) [28]:

$$\begin{aligned} \underset{\mathbf{W}}{\text{maximize}} \quad & \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{R}\mathbf{W}\right) \\ \text{subject to} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{D}\mathbf{W} = \mathbf{I}_K, \end{aligned} \tag{4}$$

with

$$\mathbf{W} = \begin{bmatrix}\mathbf{W}_x\\ \mathbf{W}_y\end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix}\mathbf{R}_{xx} & \mathbf{R}_{xy}\\ \mathbf{R}_{yx} & \mathbf{R}_{yy}\end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix}\mathbf{R}_{xx} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{yy}\end{bmatrix}. \tag{5}$$

From the Karush–Kuhn–Tucker (KKT) conditions, we have the following set of equations:

$$\mathbf{R}\mathbf{W} = \mathbf{D}\mathbf{W}\boldsymbol{\Lambda}, \tag{6a}$$

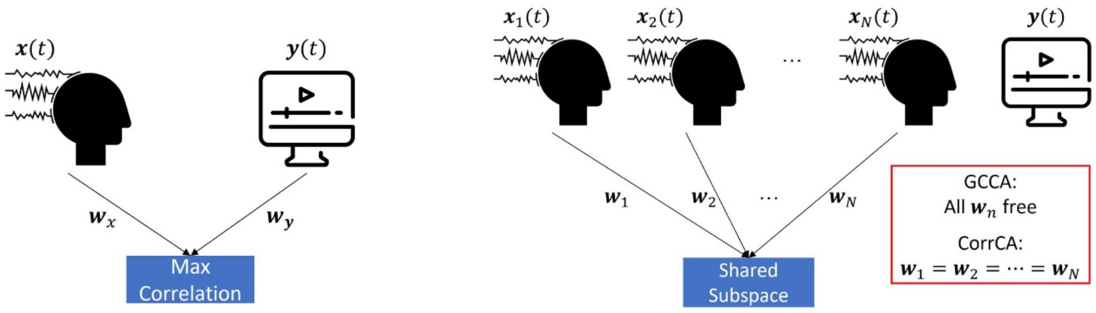
$$\mathbf{W}^{\mathrm{T}}\mathbf{D}\mathbf{W} = \mathbf{I}_K, \tag{6b}$$

where $\boldsymbol{\Lambda}$ is a symmetric matrix containing the Lagrange multipliers. By left multiplying $\mathbf{W}^{\mathrm{T}}$ to both sides of (6*a*) and making use of condition (6*b*), it is obvious that maximizing the objective function corresponds to maximizing $\mathrm{Tr}(\boldsymbol{\Lambda})$. Therefore, an optimal $\mathbf{W}$ can be obtained by solving (6*a*) as a generalized eigenvalue decomposition (GEVD) problem, and the columns of $\mathbf{W}$ are the generalized eigenvectors (GEVC) corresponding to the top-K largest generalized eigenvalues (GEVL).

3.3. Multi-subject EEG analysis

3.3.1. GCCA

Note that CCA only works for two views (in our case $\mathbf{x}(t)$ and $\mathbf{y}(t)$). When analyzing the correlation between more than two views, e.g. jointly measuring the correlation between EEG signals $\mathbf{X}_n$ of multiple subjects, it becomes necessary to extend CCA



(a) CCA

(b) (MAXVAR-)GCCA and CorrCA

Figure 2. Conceptual illustrations of stimulus-aware video-EEG analysis using CCA and multi-subject EEG analysis using (MAXVAR-)GCCA and CorrCA.

to accommodate multiple matrices as inputs. The objective is now to find the per-subject filters $\mathbf{W}_n \in \mathbb{R}^{C\times K}$ (on the EEG) such that the outputs are on average maximally correlated. The stimulus is, in this analysis, not explicitly taken into account. However, the generalization of CCA is not unique and in this work we choose MAXVAR-GCCA [29], which aims to find linear transformations for different views such that the transformed signals, on average, closely approximate an unknown shared subspace $\mathbf{S} \in \mathbb{R}^{T\times K}$. Mathematically, it is formulated as:

$$\begin{aligned} \underset{\mathbf{W}_1,\dots,\mathbf{W}_N,\mathbf{S}}{\text{minimize}} \quad & \sum_{n=1}^{N} \|\mathbf{S}-\mathbf{X}_n\mathbf{W}_n\|_{\mathbb{F}}^2 \\ \text{subject to} \quad & \mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_K, \end{aligned} \tag{7}$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm. Using the KKT conditions, we can show that the solution to (7) is also given by a GEVD problem that has a similar form as (6a), but with a different content in the block matrices [30]:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_N \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{N1} & \cdots & \mathbf{R}_{NN} \end{bmatrix},$$
$$\mathbf{D} = \begin{bmatrix} \mathbf{R}_{11} & 0 & \cdots & 0 \\ 0 & \mathbf{R}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R}_{NN} \end{bmatrix}, \tag{8}$$

with $\mathbf{R}_{ij} = \frac{1}{T}\mathbf{X}^{\mathrm{T}}{}_i\mathbf{X}_j$, the cross-correlation matrix between the EEGs of the two subjects when $i \neq j$ and the autocorrelation matrix of subject i when $i = j$. An optimal $\mathbf{W}$ is the horizontal stack of the GEVCs corresponding to the K largest GEVLs. Therefore, up to a scaling factor in the solution, MAXVAR-GCCA is equivalent to (4) with parameters defined as in (8), which is a natural extension of CCA. The shared subspace can be obtained directly from the KKT conditions as

$$\mathbf{S} = \sum_{n=1}^{N} \mathbf{X}_n\mathbf{W}_n\mathbf{\Lambda}^{-1}. \tag{9}$$

3.3.2. CorrCA

CorrCA was proposed in [17], also for quantifying the correlation between the neural data of multiple subjects. A key difference between CorrCA and GCCA is that CorrCA constrains the filters of different subjects to be the same. A multi-component formulation of CorrCA is:

$$\begin{aligned} \underset{\mathbf{V}_s}{\text{maximize}} \quad & \sum_{i=1,i\neq j}^{N}\sum_{j=1}^{N} \mathrm{Tr}\left(\mathbf{V}^{\mathrm{T}}{}_s\mathbf{R}_{ij}\mathbf{V}_s\right) \\ \text{subject to} \quad & \sum_{i=1}^{N} \mathbf{V}^{\mathrm{T}}{}_s\mathbf{R}_{ii}\mathbf{V}_s = \mathbf{I}_K, \end{aligned} \tag{10}$$

where the columns of $\mathbf{V}_s \in \mathbb{R}^{C\times K}$ are the shared filters. From this definition, it is clear that (10) can be viewed as a straightforward multi-view extension of the (2-view) CCA in (4) with an extra constraint $\mathbf{W}_1 = \ldots = \mathbf{W}_N = \mathbf{V}_s$. One can show that CorrCA can also be viewed as a MAXVAR-GCCA with an additional constraint that all data views share the same filter (compare with (7)):

$$\begin{aligned} \underset{\mathbf{V}_s,\mathbf{S}}{\text{minimize}} \quad & \sum_{n=1}^{N} \|\mathbf{S}-\mathbf{X}_n\mathbf{V}_s\|_{\mathbb{F}}^2 \\ \text{subject to} \quad & \mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_K, \end{aligned} \tag{11}$$

which is more insightful since it also gives a well-defined shared subspace. Equations (10) and (11) are equivalent (up to a scaling factor), and their optimal solutions can be obtained by horizontally stacking the GEVCs corresponding to the K largest GEVLs of the following GEVD problem (appendix B):

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{R}_{ij}\right)\mathbf{V}_s = \left(\sum_{i=1}^{N}\mathbf{R}_{ii}\right)\mathbf{V}_s\mathbf{\Lambda}. \tag{12}$$

A conceptual illustration of MAXVAR-GCCA and CorrCA can be found in figure 2(b). The constraint that the filters of different views should be the same reduces the parameters in the model and thus mitigates the overfitting problem when there is insufficient data. However, this constraint also imposes limitations on the applicability of the model, as it requires the views to have the same dimension (e.g. same number of EEG channels per subject). In comparison, GCCA allows different views to have different dimensions, so it can potentially be used to jointly analyze correlations across different data modalities or, for example, when a different number of channels or setup is used per subject. In addition, GCCA can tolerate misalignments between the data of different subjects to some extent by additionally using temporal filters. CorrCA, however, requires exact temporal alignment of the different signals for an optimal performance, even with spatial-temporal filters.

3.4. Filter design

In this study, we used spatial-temporal filters by default. If the data is one-dimensional, then they automatically reduce to temporal filters. For CCA, the filters applied to the video feature(s) had $L_y = 15$ lags, capturing information from the preceding 500 ms of video content. We included $L_x = 3$ lags in the filters of EEG signals, encompassing information from the previous, current, and next sample (i.e. from −33.3 ms to 33.3 ms). In GCCA or CorrCA, the filters had $L_x = 5$ lags, ranging from the past two samples to the next two samples (i.e. from −66.6 ms to 66.6 ms).

3.5. Evaluation metrics

We used 10-fold cross validation to evaluate the performance of the algorithms. Each dataset was divided into 10 folds, with each fold used as the test set once while the remaining folds served as the training set. The results on the test sets were averaged across 10 folds. In the stimulus-aware video-EEG analysis, we applied the filters obtained using CCA in the training stage to the test set. For each canonical component pair k, we computed the Pearson correlation coefficient ρ_k between the transformed features and the transformed EEG signals. In addition to analyzing the results of individual canonical components, it is also valuable to evaluate the performance of multiple canonical components collectively. To achieve this, we utilized the total squared correlation (TSC) metric proposed in [31]:

$$\Theta = \sum_{k=1}^{\min\left\{C,D_y\right\}} \rho_k^2, \tag{13}$$

which is related to the proximity of the EEG and video feature spaces. A higher value indicates a closer correspondence. However, in this study, we were less interested in the distance between the original EEG and video feature spaces, as it may be heavily affected by noise. Therefore, we used a slightly modified version of the TSC metric to consider only the first K ($K = 2$ in the video-EEG analysis) canonical component pairs that were less affected by noise:

$$\Theta = \sum_{k=1}^{K} \rho_k^2. \tag{14}$$

In the multi-subject EEG analysis, we applied GCCA or CorrCA filters to transform the EEG signals of each subject. We then calculated the inter-subject correlation (ISC) for each component by averaging the pairwise correlations between the transformed EEG signals of two subjects, considering all possible subject combinations [17]. Similar to the stimulus-aware video-EEG analysis, we used the TSC metric ($K = 4$) to jointly consider multiple canonical components of two subjects. By averaging the TSC values across all subject pairs, we obtained the inter-subject total squared correlation (ISTSC).

To assess the significance of the results, we employed a permutation test. The null hypothesis assumed that the transformed data views were uncorrelated. To simulate this scenario, we first obtained the transformed data with the trained filters, and then created the permuted copies by randomly shuffling the transformed EEG samples (or/and the transformed features) 1000 times. The p-value was determined as the proportion of absolute correlation coefficients (or ISC values) calculated from the permuted data that exceeded the correlation between the original transformed data. If the p-value was below the α-level (set to 0.05), we rejected the null hypothesis and concluded that the correlation was statistically significant.

For comparing the performance of different features, we employed the paired Wilcoxon signed-rank test, a non-parametric statistical hypothesis test for comparing paired observations. In our study, pairs could be, e.g. a pair of TSCs obtained with two different features for a subject. The null hypothesis assumed that the median difference between the pairs of observations was zero. In two-sided tests, the alternative hypothesis was that the median difference was not equal to zero. We primarily used one-sided tests in this study, where the alternative hypothesis was that the median difference was greater (lower) than zero. We applied a Bonferroni correction when multiple comparisons were performed.

The quality of the features can also be evaluated based on their performance in a specific task. A common practice involves using these features in a MM task [13, 32]. An illustration of the MM task is shown in figure 3. In this study, we trained the filters of the EEG signals and video features using CCA, and divided the test set into N_s 1-min EEG and

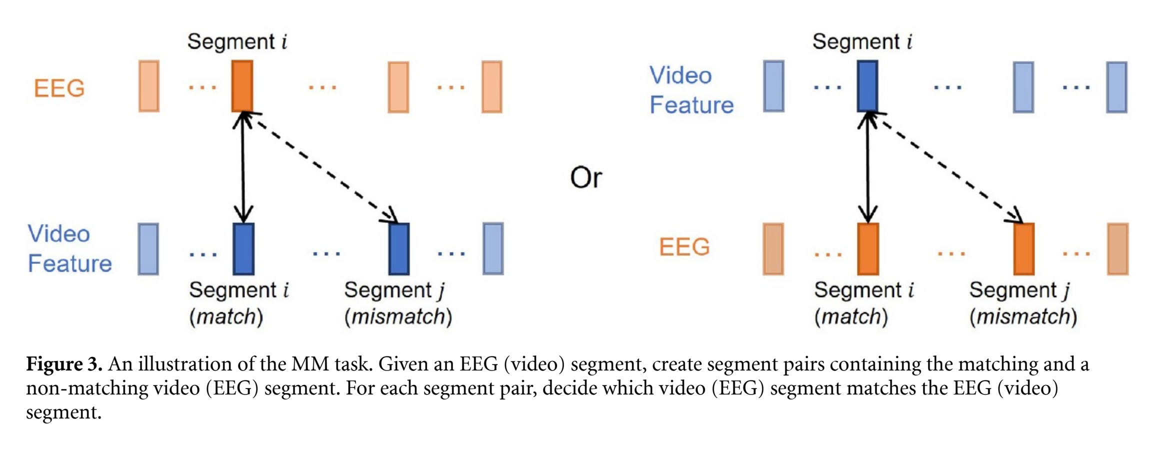


Figure 3. An illustration of the MM task. Given an EEG (video) segment, create segment pairs containing the matching and a non-matching video (EEG) segment. For each segment pair, decide which video (EEG) segment matches the EEG (video) segment.

video segments. For each EEG segment in the test set, there were $(N_s - 1)$ test pairs, where each pair consisted of the matching video segment combined with a non-matching segment. We applied filters obtained in the training stage and computed the correlations between the transformed EEG segment and the two transformed video segments, respectively. The video segment with a higher correlation was selected as the match. The total number of tests conducted was $N_s(N_s - 1)$, and the error rate was calculated as the proportion of incorrect decisions. We also analyzed the dual problem, i.e. for each video segment, we created EEG segment pairs and performed the MM task. The error rate was averaged across 10 folds.

3.6. Interpretation of weights

Correlation coefficients and TSCs provide valuable insights into the degree of correlation between data views or the proximity of data spaces. However, to gain further understanding, we can also look into the weights of the filters. For temporal filters, a conventional practice to interpret the weights is to plot the frequency response, which shows the frequency components that are amplified or attenuated by the filter. Regarding spatial filters, an intuitive way is to identify the channels with the highest contribution by looking at the absolute values of the weights and then locate the regions of interest. However, as argued in [33], this approach can be misleading since channels that are not highly relevant to the extracted components may also receive large weights due to, e.g. noise cancelling. A better approach is to compute the forward models, which involves reconstructing the original EEG signals from the extracted components, and then plot the topographic maps of the weights of the forward models (for both the video-EEG and multi-subject EEG analysis). The extracted components are better reflected in channels with larger (unsigned) weights. As the unrelated EEG background noise can, in theory, not be predicted from these components, these forward models purely reflect the stimulus-related contributions and not the noise.

For CCA, following the notations in section 3.2 and adding subscripts to matrices to specify the subject, we define the transformed individual EEG signal as $\mathbf{S}_n = \mathbf{X}_n\mathbf{W}_n \in \mathbb{R}^{T\times K}$ and the individual forward model as $\mathbf{F}_n \in \mathbb{R}^{C\times K}$ for subject n. The forward model can be computed by solving the following LS problem:

$$\min_{\mathbf{F}_n} \|\mathbf{X}_n - \mathbf{S}_n{\mathbf{F}^{\mathrm{T}}}_n\|_{\mathbb{F}}^2. \tag{15}$$

The solution is

$$\mathbf{F}_n = {\mathbf{X}^{\mathrm{T}}}_n\mathbf{S}_n\left({\mathbf{S}^{\mathrm{T}}}_n\mathbf{S}_n\right)^{-1} = {\mathbf{X}^{\mathrm{T}}}_n\mathbf{X}_n\mathbf{W}_n\left({\mathbf{W}^{\mathrm{T}}}_n{\mathbf{X}^{\mathrm{T}}}_n\mathbf{X}_n\mathbf{W}_n\right)^{-1}. \tag{16}$$

For GCCA (and CorrCA), we reconstruct signals from the shared subspace $\mathbf{S}$ and define an averaged version of the forward model $\mathbf{F}$ as

$$\min_{\mathbf{F}} \left\| \begin{bmatrix}\mathbf{X}_1\\ \vdots\\ \mathbf{X}_N\end{bmatrix} - \begin{bmatrix}\mathbf{S}\\ \vdots\\ \mathbf{S}\end{bmatrix}\mathbf{F}^{\mathrm{T}} \right\|_{\mathbb{F}}^2. \tag{17}$$

Making use of the constraint that $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_K$ and a KKT condition ${\mathbf{X}^{\mathrm{T}}}_n\mathbf{S} = {\mathbf{X}^{\mathrm{T}}}_n\mathbf{X}_n\mathbf{W}_n$ of (7), the solution can be obtained as:

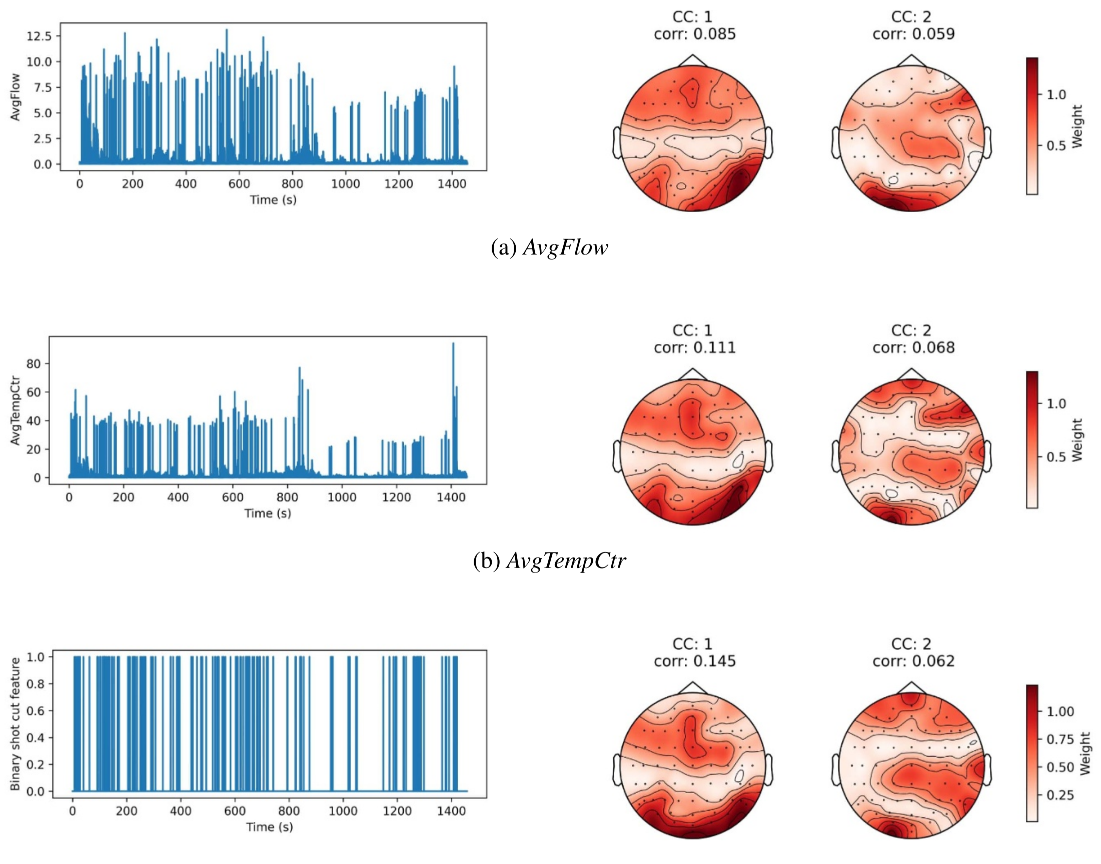
$$\mathbf{F} = \frac{1}{N}\sum_{n=1}^{N}{\mathbf{X}^{\mathrm{T}}}_n\mathbf{X}_n\mathbf{W}_n = \frac{T}{N}\mathbf{D}\mathbf{W}. \tag{18}$$

Note that when using spatial-temporal filters, the form of the forward models will be slightly different from (16) and (18) since $\mathbf{S}$ (and $\mathbf{S}_n$) will be computed from a lagged version of $\mathbf{X}_n$. For each component, we can generate a topographic plot illustrating the weights of the forward model. To ensure clarity in visualization, we always plot the absolute values of the weights.

4. Results

4.1. Mind shot cuts

Using the MrBean dataset, we first show that the shot cuts in the videos lead to prominent peaks in *AvgFlow*



(a) *AvgFlow*

(b) *AvgTempCtr*

(c) Binary shot cut feature

Figure 4. Comparison of CCA between EEG and video stimulus using *AvgFlow*, *AvgTempCtr*, and the binary shot cut feature for a representative subject in the MrBean dataset. The features are presented on the left of each subfigure, and the corresponding forward models and correlation coefficients of the first two canonical components are plotted on the right. All plotted components are significant. With the binary shot cut feature, the correlation coefficients are higher or comparable to those obtained by CCA with *AvgFlow* and *AvgTempCtr*. The forward models also share similar patterns when using the three different features.

and *AvgTempCtr*, which actually dominate the correlations found by CCA.

An underlying assumption in Gunnar-Farneback Optical Flow (section 3.1.1) is that the objects in a video sequence tend to move coherently and smoothly, which is clearly violated when there is a shot cut. Most of the pixels in the previous frame will have large displacement or even have disappeared in the next frame, resulting in a spurious peak in *AvgFlow* (figure 4(a)). Similarly, shot cuts usually lead to large intensity changes, resulting in similar peaks in *AvgTempCtr* (figure 4(b)). To automatically identify these peaks, we used a peak detection algorithm (find_peaks [34]) on the *AvgFlow* (or *AvgTempCtr*) feature, yielding 124 (or 125) prominent peaks. To verify that these peaks correspond to shot cuts, we applied a shot change detection method (AdaptiveDetector() from the PySceneDetect library [35]) to the video, which returned 127 time points of the shot cuts. Out of the 124 peaks in *AvgFlow* (and 125 peaks in *AvgTempCtr*), 122 (and 124, respectively) were matched with the shot cuts.

Since the amplitude of the peaks caused by shot cuts no longer indicates the magnitude of velocity or the temporal intensity change, these peaks should be treated as artifacts. Therefore, apart from applying CCA to the original features, we explored a second setting where we first identified shot cuts based on the results of the shot change detection function. Subsequently, we removed one second of data before and after each shot cut, and performed CCA again. This procedure resulted in a loss of approximately 4 min of data, which accounted for less than 20% of the entire dataset, and allows to probe the correlations with *AvgFlow* and *AvgTempCtr* when no shot cuts are present. Thirdly, instead of removing shot cuts, it is also interesting to see what would happen if we exclusively retain shot cut information. For this purpose, we designed a binary shot cut feature that assigned a value of 1 if a shot cut was present in the corresponding frame, and 0 otherwise (figure 4(c)). In this third setting, we used CCA to correlate the binary shot cut feature with EEG signals. We present topographic plots of the forward models in three settings for a representative subject in figure 4. The resulting correlations for the first two canonical components, under three settings, for all subjects are shown in figure 5.

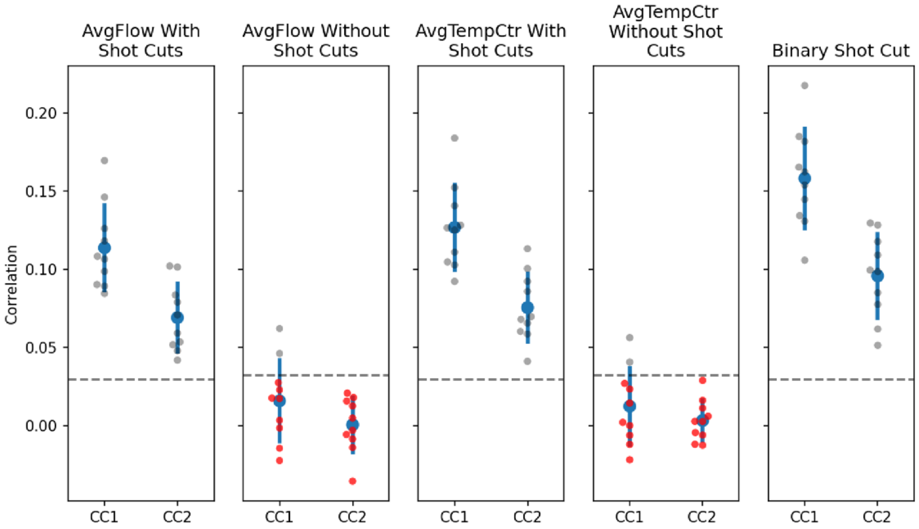


Figure 5. The per-subject Pearson correlation coefficients of the first two canonical components (CC) obtained on the MrBean dataset with CCA under three settings: 1) correlating original *AvgFlow* (*AvgTempCtr*) features with EEG signals; 2) correlating *AvgFlow* (*AvgTempCtr*) features with EEG signals when shot cuts are removed (both from EEG and the features); 3) correlating binary shot cut features with the EEG signals. The dots represent per-subject correlations, and the bars indicate the mean and standard deviation of these correlations. The horizontal dashed lines show the significance level and non-significant correlations are marked in red.

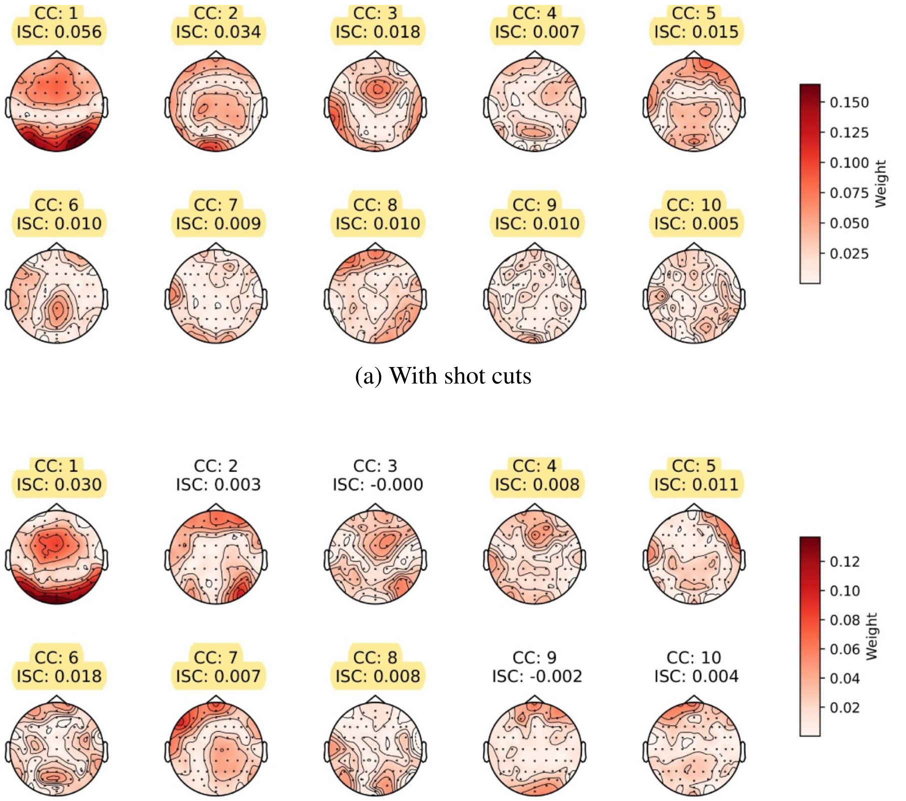
We can observe in the topographic plots (figure 4) that the forward models of the canonical components obtained with *AvgFlow*, *AvgTempCtr*, and the binary shot cut feature are similar to each other. In figure 5, by comparing the results of the first two settings, we can see that the correlations drop drastically and even become non-significant after removing the shot cuts, both for *AvgFlow* and *AvgTempCtr*. In contrast, the correlations obtained with the binary shot cut feature are comparable to or even higher than the ones obtained with the original *AvgFlow* or *AvgTempCtr* feature. Based on these observations, we argue that the correlations obtained with *AvgFlow* and *AvgTempCtr* using CCA are primarily driven by shot cuts. When shot cuts are removed from the natural videos, these features are inadequate in consistently finding significant correlations, suggesting their limited utility in capturing the temporal coupling between EEG signals and natural single-shot videos.

Since the binary shot cut feature exhibits a strong correlation with EEG signals, we infer that shot cuts elicit robust (probably ERP-like) neural responses in the brain. Given that the data of all subjects are synchronized, it is expected that the EEG components associated with shot cuts also show coherence across subjects and can thus be captured by GCCA/CorrCA. After removing the shot cuts, those EEG components may again disappear or change. To test this intuition, we applied CorrCA to the EEG data, both with and without shot cuts removed. CorrCA was chosen as it adds additional regularization, which is required given the relatively small scale (24 min × 10 subjects) of the dataset. Figure 6 illustrates the ISCs and forward models of the top-10 canonical components before and after the removal of shot cuts. The number of significant components decreased from 10 to 6. The vanished components may be associated with the neural responses elicited by shot cuts, which also implies that one factor can affect multiple EEG components. Among the remaining significant components, the average ISC drops by 42.7% compared to the top 6 components with highest ISC in figure 6(a), and the reduction may also be related to shot cuts.

These results demonstrate that shot cuts have a significant impact on both video-EEG analysis with the traditional *AvgFlow* and *AvgTempCtr* features and multi-subject EEG analysis. These shot cuts lead to significant correlations, which are useful as they could indicate whether a subject is watching a video or not. However, in many application scenarios, visual stimuli typically do not contain shot cuts. For instance, when determining a driver's focus on road conditions using EEG signals and video streaming from a dashcam, consecutive frames will have smooth transitions. Moreover, figure 5 clearly shows that *AvgFlow* and *AvgTempCtr* are unable to extract significant components in-between shot cuts, such that more temporally fine-grained estimations of levels of attention or engagement are impossible. To confirm these intial findings, in the next section, we will investigate the performance of *AvgFlow* and *AvgTempCtr* in single-shot videos, without any shot cuts, and propose our novel object-based features from section 3.1.2 as alternatives.

4.2. Object-based features lead to significant correlation in single-shot videos

Starting from this section, we analyze the Single-Shot dataset, with video clips without shot cuts. In table 2,



(a) With shot cuts

(b) Shot cuts removed

Figure 6. Absolute value of the weights of the forward models and ISCs obtained by CorrCA with the EEG data of 10 subjects watching Mr Bean with and without shot cuts. The significant components are highlighted.

we compare the performance of CCA using the original features *AvgFlow* and *AvgTempCtr* with their newly proposed object-based counterparts, *ObjFlow* and *ObjTempCtr*. It is evident that the use of object-based features leads to substantial improvements in robustness, i.e. the feature is able to consistently find significant correlations in all subjects. In terms of *AvgFlow*, only one subject (Subject 13) exhibits two significant components, while only 25% of the correlations across both components and all subjects is significant. In comparison, by employing the proposed *ObjFlow*, we are able to identify at least one significant component for each subject, with 85% of the correlations significant across both components and all subjects. While *AvgTempCtr* demonstrates slightly better robustness compared to *AvgFlow*, incorporating *ObjTempCtr* again clearly results in a drastic improvement.

Apart from comparing the robustness across subjects, we can also directly compare the strength of correlations. From table 2, we observe that the first canonical components do not always capture the most correlated information in practice, although they should theoretically. Indeed, sometimes the second component exhibits a higher correlation, which may be attributed to overfitting on the training set or small differences between the training and test set. Due to these inconsistencies in the ordering of the components (according to correlation values) between the training and test set, it makes more sense to consider the canonical components jointly using TSC. The results are plotted in figure 7. To compare the performance of different features, we conducted one-sided Wilcoxon signed-rank tests. The results indicate that *ObjFlow* and *ObjTempCtr* exhibit significant superiority over *AvgFlow* and *AvgTempCtr*, with p-values < 0.001 in both cases.

These results confirm the results from section 4.1, i.e. that the traditional *AvgFlow* and *AvgTempCtr* are inadequate to capture the temporal correlations between the EEG and an attended video when there are no shotcuts. Moreover, the results clearly show that *ObjFlow* and *ObjTempCtr* are better features that consistently yield significant correlations. This agrees with our intuition that the feature is more specific and emphasizes the regions where viewers are likely to focus. However, it is worth noting that *ObjFlow* may be correlated with eye movements if participants tracked the object during passive viewing. Consequently, the residual eye motion artefacts in the EEG signals could potentially drive significant correlations (despite regressing out the EOG signals from the EEG data), posing challenges in determining whether higher-level cognitive processes, such as movement perception, contribute to these correlations. A similar issue may arise with *ObjTempCtr*,

Table 2. Pearson correlation coefficients of the first two canonical components (CC) obtained from the Single-Shot dataset with CCA using *AvgFlow*, *AvgTempCtr*, and their object-based versions *ObjFlow* and *ObjTempCtr*. The significant correlations are in bold.

	AvgFlow		*ObjFlow*		*AvgTempCtr*		*ObjTempCtr*	
Subject	CC1	CC2	CC1	CC2	CC1	CC2	CC1	CC2
1	**0.026**	0.015	**0.034**	**0.024**	**0.027**	−0.011	**0.039**	**0.026**
2	0.013	0.010	**0.062**	**0.019**	0.015	0.006	**0.068**	**0.023**
3	**0.025**	0.011	0.016	**0.040**	**0.024**	0.016	**0.019**	**0.030**
4	0.013	0.012	**0.069**	0.016	0.012	0.010	**0.067**	**0.023**
5	**0.024**	0.016	**0.077**	**0.022**	**0.047**	0.004	**0.073**	**0.027**
6	0.000	**0.027**	**0.057**	**0.042**	**0.024**	**0.022**	**0.050**	**0.030**
7	−0.004	**0.021**	**0.033**	**0.054**	−0.004	**0.023**	**0.035**	**0.045**
8	0.009	0.001	**0.032**	0.008	**0.024**	−0.015	**0.036**	0.004
9	**0.038**	0.018	**0.085**	**0.045**	**0.043**	**0.019**	**0.076**	**0.043**
10	0.013	**0.026**	**0.061**	**0.030**	**0.028**	**0.027**	**0.063**	**0.031**
11	−0.002	0.014	**0.050**	0.015	0.015	0.011	**0.049**	**0.027**
12	−0.003	−0.007	**0.025**	0.007	−0.004	0.003	0.017	0.009
13	**0.023**	**0.025**	**0.034**	**0.022**	**0.027**	0.004	**0.025**	0.009
14	0.013	0.012	0.002	**0.021**	−0.001	**0.020**	0.001	0.015
15	0.009	0.003	**0.066**	**0.032**	**0.047**	0.000	**0.069**	**0.028**
16	−0.001	0.002	**0.039**	**0.026**	0.012	0.002	**0.039**	**0.021**
17	0.014	0.013	**0.060**	**0.048**	**0.023**	**0.023**	**0.058**	**0.055**
18	0.015	**0.031**	**0.023**	**0.027**	0.004	**0.022**	0.016	**0.038**
19	0.006	0.015	**0.047**	**0.020**	**0.032**	0.011	**0.055**	**0.030**
20	0.017	0.018	**0.047**	**0.027**	0.016	0.014	**0.058**	**0.023**

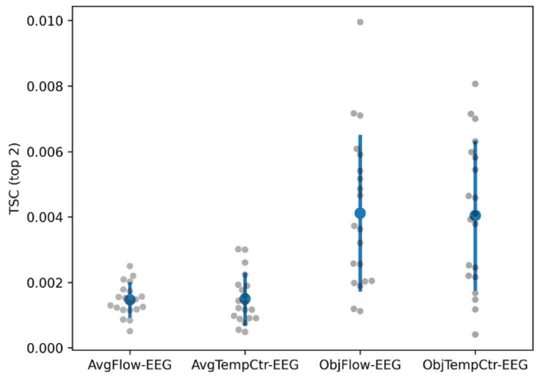


Figure 7. The per-subject individual TSCs (denoted by grey dots) between the EEG signal and *AvgFlow*, *ObjFlow*, *AvgTempCtr*, and *ObjTempCtr* in the Single-Shot dataset. The bars indicate the mean and standard deviation of the TSCs across subjects.

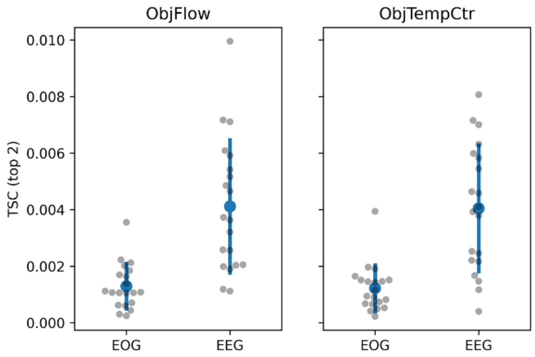


Figure 8. The per-subject individual TSCs (denoted by grey dots) when correlating the proposed object-based features with both EEG and EOG signals in the single-shot dataset. The bars indicate the mean and standard deviation of the TSCs across subjects.

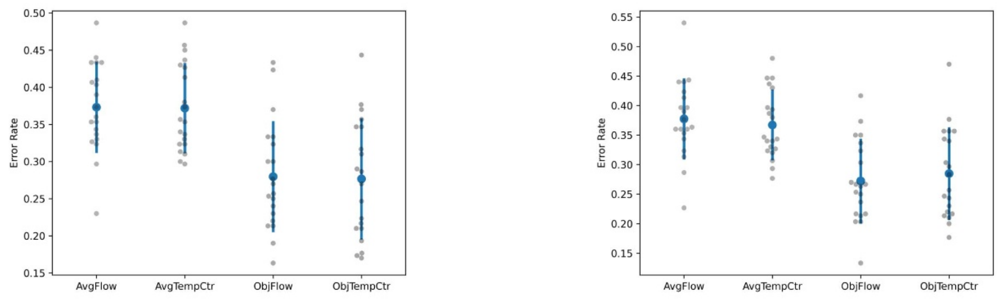
as it implicitly encodes object motion information. Therefore, we will further investigate whether the eye movements drive these correlations in the following section.

4.3. Are correlations driven by eye movements?

To check whether the correlations are driven by eye movements, we correlated the EOG signals with the proposed object-based features and compared the results with those obtained from EEG signals in the Single-Shot data set. We observed a significant decrease in TSCs of *ObjFlow*-EOG and *ObjTempCtr*-EOG (figure 8), with p-values < 0.001. This suggests that the leakage of EOG signals into the EEG cannot fully explain the correlations between the EEG signals and our features; instead, neural activities captured by EEG dominate the correlations. We conclude that eye movements do not dominantly drive the correlations.

4.4. Object-based features perform better in MM task

In section 4.2, we have demonstrated that object-based features exhibit higher correlations with EEG signals compared to traditional features. To further demonstrate the superiority of these object-based features in a more quantitatively verifiable setting, we conducted the MM task on the Single-Shot dataset using *AvgFlow*, *ObjFlow*, *AvgTempCtr*, and *ObjTempCtr*, respectively, as described in section 3.5. We retained the top 5 canonical components obtained with CCA in the training set. In the decision-making phase, we selected the one with highest correlation for



(a) Given EEG segments, select corresponding video segments.

(b) Given video segments, select corresponding EEG segments.

Figure 9. The per-subject individual error rates of the MM task (denoted by grey dots) using *AvgFlow*, *ObjFlow*, *AvgTempCtr*, and *ObjTempCtr* in the Single-Shot dataset. The bars indicate the mean and standard deviation of the error rates across subjects.

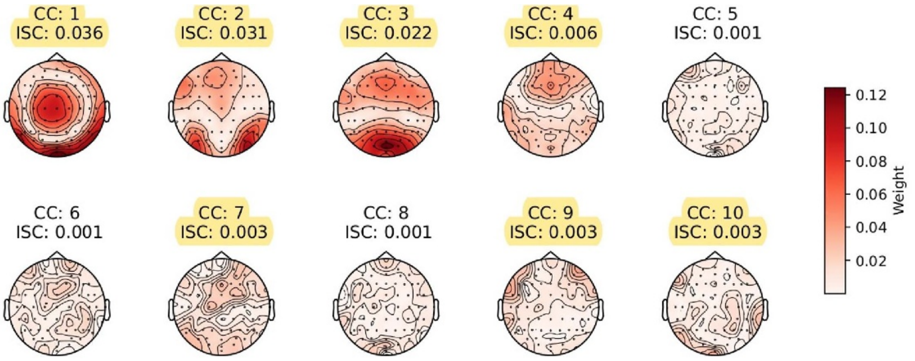


Figure 10. ISCs and forward models of the top 10 EEG components obtained by GCCA in the Single-Shot dataset. The significant components are highlighted.

each EEG-video pair, and from each pair we selected the segment with the highest correlation as the 'matched' segment. The error rates are presented in figure 9. In both scenarios, i.e. matching video segments given EEG segments and matching EEG segments given video segments, object-based features yielded lower error rates compared to their traditional counterparts, with p-values < 0.001. This further supports the argument that the proposed object-based features are more effective in identifying meaningful temporal correlations between EEG and video signals.

4.5. Multi-subject EEG analysis on single-shot videos

One limitation of doing stimuli-response analysis using CCA is that it can only be performed individually and thus cannot leverage information across subjects. With multi-subject EEG analysis, we can extract the shared subspace of EEG signals of all the subjects, which is spanned by the coherent EEG components that are time-locked to the video stimuli. These coherent EEG components have higher SNR since the asynchronous noise and background EEG activities are suppressed. Therefore, it is informative to show their forward models, which can reveal the regions where these coherent EEG components are more reflected. Given that the Single-Shot dataset is a larger dataset (63 min × 20 subjects) than the MrBean dataset (24 min × 10 subjects) used in section 4.1, we chose GCCA instead of CorrCA for the multi-subject analysis. The ISCs and the forward models (defined in (17)) of the first 10 canonical components are shown in figure 10, and the ISTSC of the top 4 canonical components is 0.0066.

The darker parts in the topographic plots indicate the regions where the coherent (across subjects) EEG components are more prominently reflected. Intuitively, the channels in those regions may be more relevant and contribute more to identifying correlations across subjects. To validate that hypothesis, we ranked the channels based on the absolute values of the forward model of the first GCCA component. We then performed a greedy forward channel selection, starting with the highest-ranked channel and iteratively adding more channels. For each new set of channels, we reran GCCA and obtained the ISC of the first component corresponding to that specific number of channels. For any number of channels, a random selection is performed 30 times, allowing to estimate

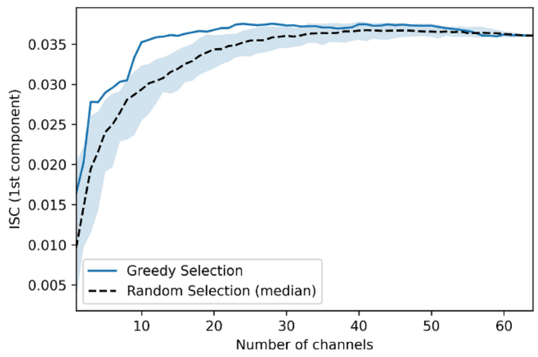


Figure 11. Evolution of ISCs of the first component as the number of selected channels increases using the Single-Shot dataset. The solid blue line corresponds to the results of greedy forward channel selection, where channels were greedily selected based on the forward model of the first GCCA component. For comparison, the dashed line shows the median of different runs where channels were randomly selected. The shaded area indicates the 5% and 95% quantiles of the random selection.

a distribution. The ISCs of the first component versus the number of channels are plotted in figure 11. In the case of greedy selection (the solid blue line), using the top 10 channels boosts the ISC of the first component to a level comparable to using all 64 channels. With random selection (median indicated by the dashed line), the trend is similar, but the ISCs are almost consistently lower when using fewer than 32 channels. We therefore conclude that the weights of the forward models are good indicators of the relevance of the corresponding channels. In this experiment, it appears that the elicited neural responses are most prominent in the occipital-temporal region.

4.6. Proportion of variance explained

A follow-up question arising from the multi-subject EEG analysis is whether we can isolate a subset of the obtained coherent EEG components that are dominantly driven by our features. This can be achieved qualitatively by first regressing out the features from the EEG signals (for the entire dataset), then reapplying the GCCA algorithm and identifying the components that either disappear or exhibit substantial changes. Results of GCCA with *ObjFlow* regressed out from EEG signals can be found in figure 12. Notably, these results closely resemble those in figure 10, particularly for the first 3 components with higher ISCs, suggesting that *ObjFlow* may not be the dominant feature for any of them. Since *ObjTempCtr* is highly correlated with *ObjFlow* in our dataset (indicated by a correlation coefficient of 0.857), the forward models when *ObjTempCtr* is regressed out are similar and we omit them for brevity.

To quantitatively assess the extent to which coherent EEG components obtained by GCCA can be attributed to our features, we can calculate the proportion of variance in the coherent stimulus responses explained by *ObjFlow* (or *ObjTempCtr*). Specifically, the variance of the stimulus responses in the k-th coherent EEG component can be estimated by the averaged pairwise covariance of the transformed EEG signals, i.e. $\frac{N(N-1)}{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathrm{Cov}(\mathbf{X}_i\mathbf{w}_{i,k},\mathbf{X}_j\mathbf{w}_{j,k})$, where $\mathbf{w}_{i,k}$ denotes the k-th column of $\mathbf{W}_i$. We refer to this quantity as inter-subject covariance (ISCOV). Since the (incoherent) background EEG components are orthogonal across subjects, they will not influence this ISCOV, except for a residue due to finite sample sizes. In order to estimate the error on the ISCOV due to this residue, we performed the following bootstrap procedure: randomly shifting the data in the test set across subjects with at least 10 s, using the pre-trained GCCA filters to transform the shifted data, and then computing the ISCOV. This procedure was repeated 100 times, with the outcomes aggregated across all components. Since the ISCOV is an averaged value across different subject pairs, the obtained ISCOVs can be modeled as a Gaussian distribution according to the central limit theorem. The 95% confidence interval of the ISCOV, obtained using the shifted EEG data and serving as the error estimate, is $(-4.4\times10^{-9},4.3\times10^{-9})$.

Table 3 presents the ISCOVs when using the original EEG data and using the EEG data with *ObjFlow* or *ObjTempCtr* regressed out. From the numbers, it appears that component 4 is most related to our features given that the ISCOV decreases by 58.7% with *ObjFlow* regressed out and by 63.5% with *ObjTempCtr* regressed out. It is also notable that the ISCOVs no longer exceed the upper limit of the 95% confidence interval for the estimation error after regression. However, despite these reductions, the forward model of component 4 does not exhibit significant changes, which might imply that our features could be coincidentally highly correlated with the actual features driving the responses in component 4.

To aggregate the effects across different components, we define the total variance as the sum of ISCOVs over the selected components. The proportion of variance explained by *ObjFlow* (or *ObjTempCtr*) is then computed as the complement of the ratio between the total variance obtained using the EEG data with *ObjFlow* (or *ObjTempCtr*) regressed out and that using the original EEG data, i.e.

$$\text{Proportion of variance explained} = 1-\frac{\sum_k \mathrm{ISCOV}_k^{\text{regressed}}}{\sum_k \mathrm{ISCOV}_k^{\text{original}}}. \tag{19}$$

For the top 4 most prominent coherent components whose ISCOVs exceed the upper limit of the 95% confidence interval for the estimation error before regression, the proportion of variance explained is 6.9% by *ObjFlow* and 7.4% by *ObjTempCtr*. In other words, around 93% of the variance in the coherent stimulus responses remains unexplained, which

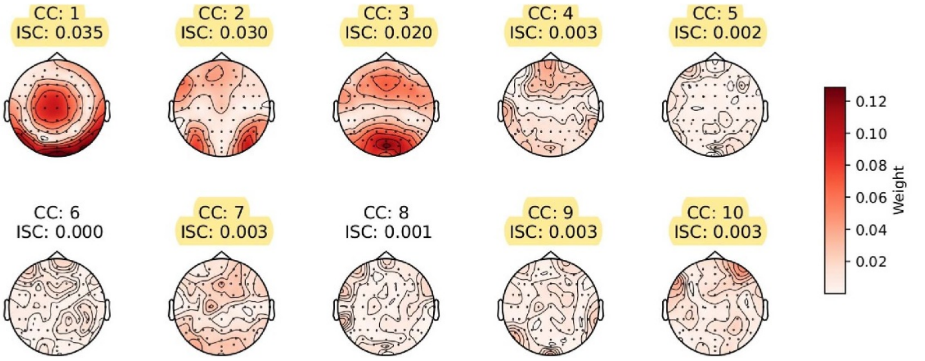


Figure 12. ISCs and forward models of top 10 EEG components obtained by GCCA with *ObjFlow* regressed out from the EEG signals in the Single-Shot dataset. The significant components are highlighted.

Table 3. ISCOVs (1×10^{-8}) of the first 10 canonical components (CC) obtained from the Single-Shot dataset with GCCA under three conditions: using (1) the original EEG data; (2) EEG data with *ObjFlow* regressed out; (3) EEG data with *ObjTempCtr* regressed out. The 95% confidence interval for the estimation error is $(-4.4 \times 10^{-9}, 4.3 \times 10^{-9})$.

	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	CC10
Original	4.32	3.64	2.33	0.63	0.09	0.13	0.31	0.10	0.28	0.33
ObjFlow regressed out	4.16	3.57	2.18	0.26	0.17	0.02	0.25	0.10	0.28	0.24
ObjTempCtr regressed out	4.11	3.56	2.21	0.23	0.16	0.03	0.22	0.11	0.28	0.24

indicates the presence of potentially more dominant yet undiscovered features.

5. Discussion

5.1. Shot cuts highly influence temporal correlations

In [36], the authors observed that the ISC based on fMRI recordings is different during the viewing of unedited and edited videos of dance performance. The unedited version represented a continuous view captured from a single camera, while the edited version consisted of concatenated shots from different cameras, therefore included shot cuts. The authors calculated ISC maps for each video and found that the two individual maps (edited vs unedited) exhibited broad overlap, but the edited version showed more significant voxels. These findings align with our own results in section 4.1, where we discovered that in the multi-subject analysis, the correlations were stronger when shot cuts were present and decreased significantly when shot cuts were removed. Additionally, we also observed substantial overlap in the forward models of the first components when comparing the results with or without shotcuts (figure 6).

In a recent study by Nentwich *et al* [37], participants were presented with natural videos while neural responses were recorded using intacranical EEG (also known as electrocorticography) with 6328 electrodes implanted throughout the entire brain. The analysis was performed channel-wise on an average brain: they extracted the broad-band high-frequency amplitude (BHA) ranging from 70 to 150 Hz and modeled the BHA of each channel as a convolution of the visual stimulus and an unknown TRF, which can be estimated using LS. While the primary focus of their research was on investigating the effects of semantic changes, there was a finding relevant to our study: they observed that a greater number of channels responded to film cuts compared to visual motion calculated using optical flow. This finding supports our conclusion that shot cuts dominate correlations in video-EEG analysis especially when using the traditional, non-object based versions of the optical flow and temporal contrast.

In earlier EEG studies with natural video footage as stimuli such as [17, 20, 21], the shot cuts were not removed and their effect was not the focus. Therefore, it is likely that the correlations found in these studies were almost exclusively driven by shot cuts. In [20], it was noted that the highest and the most sustained ISC coincided with the video segment having the most scene changes in the example video clip, which aligns with our argument. While in certain cases, e.g. using the correlation as a marker of engagement [17], the origins of the correlations may appear less important, it is advisable to be aware of the impact of shot cuts since they elicit strong neural responses that could potentially overshadow more

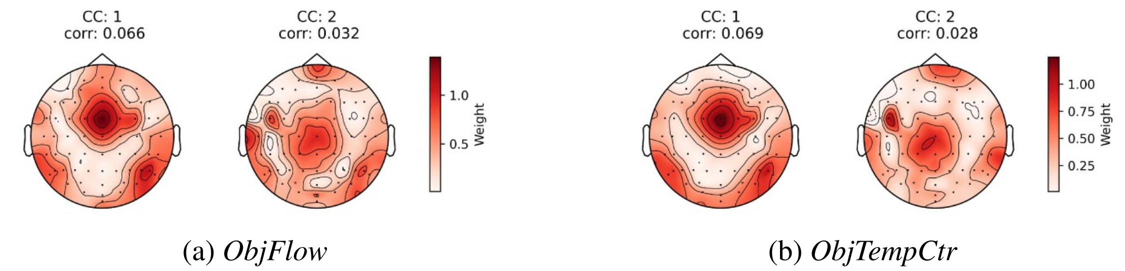


Figure 13. Comparison of forward models obtained using CCA with *ObjFlow* and *ObjTempCtr* for a representative subject in the Single-Shot dataset. For these two significant components, the forward models are highly similar.

intricate responses related to higher-level cognitive processes.

5.2. Interpretation and relations of optical flow and temporal contrast

In our experiments, both *ObjFlow* and *ObjTempCtr* lead to higher correlations with the EEG signals and lower error rates in the MM task. However, there is no significant difference in the performance of these two features. The p-value of the two-sided paired Wilcoxon signed-rank test on the TSCs obtained using *ObjFlow* and *ObjTempCtr* yields 0.90. The error rates when matching video segments given EEG segments and matching EEG segments given video segments are also not significantly different, with p-values of 0.14 and 0.59, respectively. Notably, the forward models exhibit remarkable similarity for these two features, as shown in figure 13 for a representative subject. Furthermore, stacking the two features together as a two-dimensional time series and inputting it into CCA does not show any significant difference either. These outcomes are probably attributed to the high correlation between *ObjFlow* and *ObjTempCtr* within our video dataset, as indicated by a correlation coefficient of 0.857. The high correlation comes as a surprise since *ObjFlow* and *ObjTempCtr* seem to be unrelated, representing motion and intensity changes, respectively. However, one can show that the two features are implicitly coupled, posing challenges in distinguishing their individual effects on the EEG signals (appendix C).

Therefore, based solely on the results obtained from the current dataset, we cannot conclusively determine the driving factor behind the observed correlations. While this issue is not the primary focus of our paper, we highlight it to emphasize the nontrivial nature of identifying the underlying causes of correlations when using natural videos. One feature could encode information of features with very different physical meanings, and it is difficult to disentangle their effects using uncontrolled visual stimuli. Additionally, if one feature works well in certain videos but not in others, it suggests that the feature may coincidentally correlate with the true feature that elicits the neural responses. Therefore, the diversity of videos in the training set is important for the correct interpretation of features.

5.3. Interpretation of neural patterns

Research on human movement perception also provides valuable insights into the interpretation of our results, particularly regarding the activation patterns observed in certain brain areas. For example, Grosbras *et al* conducted a meta-analysis combining fMRI results from multiple studies focused on three categories of motion: face, hands, and whole body movements [38]. They applied the activation likelihood estimation method with random effect analysis to generate a probability map reflecting the likelihood that a particular voxel was activated. Their findings revealed convergence of brain activation in the occipito-temporal and fronto-parietal regions across all categories, although with different peak locations and extents. In [39], dancers were asked to perform specific movements while detailed movement features were gathered using accelerometers. Videos of these dancers were shown to fMRI participants, and the collected data were analyzed. The researchers discovered that low-level features such as acceleration corresponded to brain regions associated with early visual and motion-sensitive areas. On the other hand, mid-level features such as dynamic symmetry mapped to the occipito-temporal cortex, posterior superior temporal sulcus, and superior parietal lobe. These findings could provide an explanation on why, in our forward models obtained with ObjFlow using CCA (e.g. figure 13(a)), the occipito-temporal and fronto-parietal regions exhibited higher levels of activation compared to other areas. However, as discussed in section 5.2, the activation of certain regions may also be related to the perception of brightness. It was found in [40] that the neural responses in the striate cortex explicitly encode brightness changes, which could be an additional explanation for the activation of the occipital area.

5.4. Potential usage of object-based features

Object-based features provide more refined representations, leading to higher and more reliable correlations with the EEG signals. These correlations could be employed as a metric for overall attention levels or engagement [22]. While our current focus is on single-object videos, in more intricate scenes involving multiple objects interacting, the proposed object-based features may still prove useful for decoding visual attention. Indeed, in [41], the results indicated that participants' selective attention mechanisms operated efficiently, being able to isolate and focus on the object of their choice despite the presence of multiple objects embedded within complex backgrounds in each scene. Therefore, by correlating features extracted from different objects with the EEG signals, it may be possible to determine the object of interest and detect shifts in visual attention. An expected constraint is that the signatures of different objects should be sufficiently distinct, otherwise the results for different objects may be too similar. Apart from decoding the attention towards a specific object, it is also desirable to measure the overall level of attention, potentially also in a multi-object setting. However, fusing the features extracted from different objects is a challenging problem that requires further investigation. For example, it is expected that the decoding process in EEG analysis is intricately influenced by the holistic interaction among diverse objects in the scene, as well as the attention and gaze direction of the subject. Exploring these confounding factors and variables becomes imperative for a comprehensive understanding when deciphering complex scenes from EEG data.

5.5. Quest for novel video features

Observing figure 6, we noted a decrease in ISCs and the number of significant components when shot cuts were removed from the MrBean dataset. Nevertheless, it is encouraging that there are still significant correlations remaining, indicating that the neural responses related to shot cuts are not the sole factors coherent across subjects. The same holds true for the Single-Shot dataset, where multiple significant components were found despite the absence of shot cuts in the videos (figure 10). This motivates the quest for novel video features that are not solely driven by shot cuts and can capture these components. The object-based features we have proposed are limited by their constraint on the number of objects in each frame. Furthermore, these features explain only approximately 7% of the variance in the coherent stimulus responses across subjects (section 4.6), suggesting that there are potentially more dominant features that are yet to be discovered. This is not particularly surprising, as both optical flow and temporal contrast are still relatively low-level features. It could be beneficial to leverage knowledge from computer vision and representation learning to investigate higher-level and more abstract features. Such features could potentially prove useful across a wide range of videos.

6. Conclusion

This study focused on identifying temporal correlations between natural video footage and EEG signals, for which two new datasets were collected. The MrBean dataset used a film clip that contains many shot cuts as the stimulus, while in the Single-Shot dataset the videos were carefully selected to be shot cut-free and to contain only a single moving object. We revealed that the correlations between video features such as optical flow and temporal contrast and the EEG signals, which were reported in previous studies, were heavily influenced by shot cuts present in the videos, leading to over-optimistic correlations between both modalities. We showed that removal of such shot cuts result in non-significant correlations in the majority of the subjects. We proposed the use of object-based features as a more robust alternative, resulting in significant correlations with the EEG signals across all subjects, even in the absence of shot cuts. Importantly, we showed that the observed correlations were not predominantly driven by eye movements, which are usually considered as confounds. Furthermore, we demonstrated that the proposed object-based features were more effective in the MM task, yielding lower error rates compared to traditional features. Finally, we illustrated that the proposed features did not dominantly drive the coherent stimulus responses, and more influential features are yet to be discovered.

For future research, there are several promising directions worth exploring. Firstly, we can shift from linear models to non-linear models to capture more complex relationships between video features and EEG signals. Secondly, exploring higher-level video features, whether with or without semantic meanings, may provide insights on how the brain processes more abstract information. Lastly, applying the proposed object-based features in a multi-object setting would reveal their effectiveness in decoding attention towards specific objects within complex visual scenes.

Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI: https://zenodo.org/records/10512414 [42].

Acknowledgments

This research is funded by the Research Foundation—Flanders (FWO) Project No. G081722N, junior postdoctoral fellowship fundamental research of the FWO (for S Geirnaert, No. 1242524 N), the

European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 802895), the Flemish Government (AI Research Program), and the PDM mandate from KU Leuven (for S Geirnaert, No PDMT1/22/009). All authors are also affiliated with Leuven.AI—KU Leuven institute for AI, Belgium.

The authors also thank the participants for their time and effort in the experiments.

Appendix A. Filters in preprocessing

In the preprocessing pipeline, we applied a high pass filter with a cutoff frequency of 0.5 Hz and a notch filter at 50 Hz. Both filters are zero phase and have a finite impulse response, designed using the window method (a Hamming window with 0.0194 passband ripple and 53 dB stopband attenuation). The specifications for each filter are as follows:

- High pass filter
  - Lower passband edge: 0.50
  - Lower transition bandwidth: 0.50 Hz (−6 dB cutoff frequency: 0.25 Hz)
  - Filter length: 6.6 s
- Notch filter
  - Lower passband edge: 49.38
  - Lower transition bandwidth: 0.50 Hz (−6 dB cutoff frequency: 49.12 Hz)
  - Upper passband edge: 50.62 Hz
  - Upper transition bandwidth: 0.50 Hz (−6 dB cutoff frequency: 50.88 Hz)
  - Filter length: 6.6 s

It is worth noting that before downsampling, a low pass filter was implicitly applied to avoid aliasing. The low pass filter, according to the documentation of the MNE-Python package [23], is a brick-wall filter applied in the frequency domain at 15 Hz (the Nyquist frequency of the desired new sampling rate).

Appendix B. Equivalence of (10) and (11)

In this section, we show that the solutions of (10) and (11) are the same up to a scaling factor. We start with the solution of (10). The Lagrange function of (10) is:

$$\mathcal{L}_1\left(\mathbf{V}_s,\mathbf{\Lambda}_1\right) = \sum_{i=1,i\neq j}^{N}\sum_{j=1}^{N}\mathrm{Tr}\left(\mathbf{V}^{\mathrm{T}}{}_s\mathbf{R}_{ij}\mathbf{V}_s\right) - \mathrm{Tr}\left(\mathbf{\Lambda}_1{}^{\mathrm{T}}\left(\sum_{i=1}^{N}\mathbf{V}^{\mathrm{T}}{}_s\mathbf{R}_{ii}\mathbf{V}_s - \mathbf{I}_K\right)\right). \tag{20}$$

The KKT conditions for $\mathbf{V}_s$ to be optimal are then given by:

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{R}_{ij}\right)\mathbf{V}_s = \left(\sum_{i=1}^{N}\mathbf{R}_{ii}\right)\mathbf{V}_s\left(\mathbf{\Lambda}_1 + \mathbf{I}_K\right), \tag{21a}$$

$$\sum_{i=1}^{N}\mathbf{V}^{\mathrm{T}}{}_s\mathbf{R}_{ii}\mathbf{V}_s = \mathbf{I}_K. \tag{21b}$$

By left multiplying (21a) by $\mathbf{V}_s{}^{\mathrm{T}}$ and using (21b), the objective function of (10) can be simplified as $\mathrm{Tr}\left(\mathbf{\Lambda}_1\right)$. Therefore, the optimal $\mathbf{V}_s$ is the horizontal concatenation of the GEVCs corresponding to the K largest GEVLs of the GEVD problem (21a) (up to orthogonal transformations). The correct scaling of the GEVCs is determined by (21b).

Similarly, for (11), we write down the Lagrangian:

$$\mathcal{L}_2\left(\mathbf{V}_s,\mathbf{S},\mathbf{\Lambda}_2\right) = \sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{S}-\mathbf{X}_n\mathbf{V}_s\right)^{\mathrm{T}}\left(\mathbf{S}-\mathbf{X}_n\mathbf{V}_s\right)\right) - \mathrm{Tr}\left(\mathbf{\Lambda}_2{}^{\mathrm{T}}\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}-\mathbf{I}_K\right)\right). \tag{22}$$

The KKT conditions for optimal $\mathbf{V}_s$ and $\mathbf{S}$ are:

$$\sum_{n=1}^{N}\mathbf{X}_n{}^{\mathrm{T}}\mathbf{S} = \left(\sum_{n=1}^{N}\mathbf{X}_n{}^{\mathrm{T}}\mathbf{X}_n\right)\mathbf{V}_s, \tag{23a}$$

$$\mathbf{S}\left(N\mathbf{I}_K - \mathbf{\Lambda}_2\right) = \sum_{n=1}^{N}\mathbf{X}_n\mathbf{V}_s, \tag{23b}$$

$$\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_K. \tag{23c}$$

Plugging (23b) into (23a) yields:

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{R}_{ij}\right)\mathbf{V}_s = \left(\sum_{i=1}^{N}\mathbf{R}_{ii}\right)\mathbf{V}_s\tilde{\mathbf{\Lambda}}_2, \tag{24}$$

with $\tilde{\mathbf{\Lambda}}_2 = (N\mathbf{I}_K - \mathbf{\Lambda}_2)$. Using (23a) and (23c), the objective function of (11) can be written as $N[K - \mathrm{Tr}(\sum_{i=1}^{N}\mathbf{V}_s{}^{\mathrm{T}}\mathbf{R}_{ii}\mathbf{V}_s)]$. From (23b) and (23c), we have $\tilde{\mathbf{\Lambda}}_2^{\mathrm{T}}\tilde{\mathbf{\Lambda}}_2 = N(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{V}_s{}^{\mathrm{T}}\mathbf{R}_{ij}\mathbf{V}_s)$. Combining it with (24) yields $N(\sum_{i=1}^{N}\mathbf{V}_s{}^{\mathrm{T}}\mathbf{R}_{ii}\mathbf{V}_s) = \tilde{\mathbf{\Lambda}}_2$. Therefore, minimizing the objective function is equivalent to maximizing $\mathrm{Tr}(\tilde{\mathbf{\Lambda}}_2)$. Then, again, the optimal $\mathbf{V}_s$ is the horizontal stack of the GEVCs corresponding to the K largest GEVLs of the GEVD problem (24). The scaling factor is determined by (23c). As (21a) and (24) represent the same GEVD problems, the solutions of (10) and (11) are identical up to a scaling factor.

Appendix C. The coupling between optical flow and temporal contrast

To illustrate the implicit coupling between *ObjFlow* and *ObjTempCtr*, we start with the calculation of velocity vectors in optical flow, which usually involves making assumptions on the pixel intensity. Take the Gunnar-Farneback Optical Flow [24] as an example. The algorithm is based on polynomial expansions, approximating the intensity $I_m(\mathbf{z})$ of some neighborhood of each pixel in the m-th frame with, e.g. a local quadratic polynomial $I_m(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{A}_m(\mathbf{z})\mathbf{z} + \mathbf{b}_m(\mathbf{z})^{\mathrm{T}}\mathbf{z} + c_m(\mathbf{z})$, where $\mathbf{z}$ denotes the two-dimensional pixel coordinate. The local parameters $\{\mathbf{A}_m(\mathbf{z}), \mathbf{b}_m(\mathbf{z}), c_m(\mathbf{z})\}$ of this model for frame m can be estimated using weighted LS. The algorithm then tries to find the displacement vector of each pixel $\mathbf{d}_m(\mathbf{z})$ under the assumption that $I_m(\mathbf{z}) = I_{m-1}(\mathbf{z} - \mathbf{d}_m(\mathbf{z}))$. By matching the coefficients of the two polynomials, we can derive $\mathbf{d}_m(\mathbf{z})$ as

$$\mathbf{d}_m(\mathbf{z}) = -\frac{1}{2}\mathbf{A}_{m-1}(\mathbf{z})^{-1}(\mathbf{b}_m(\mathbf{z}) - \mathbf{b}_{m-1}(\mathbf{z})). \quad (25)$$

The velocity vector $\mathbf{v}_m(\mathbf{z})$ can then be obtained by multiplying $\mathbf{d}_m(\mathbf{z})$ with the sampling frequency of the video f_s.

Additionally, the coefficient matching yields the following two equations: $\mathbf{A}_m(\mathbf{z}) = \mathbf{A}_{m-1}(\mathbf{z})$ and $c_m(\mathbf{z}) = c_{m-1}(\mathbf{z}) + \mathbf{d}_m(\mathbf{z})^{\mathrm{T}}\mathbf{A}_{m-1}(\mathbf{z})\mathbf{d}_m(\mathbf{z}) - \mathbf{b}_{m-1}(\mathbf{z})^{\mathrm{T}}\mathbf{d}_m(\mathbf{z})$. Since in practice $\mathbf{A}_m(\mathbf{z}) = \mathbf{A}_{m-1}(\mathbf{z})$ generally does not hold, the approximation $[\mathbf{A}_m(\mathbf{z}) + \mathbf{A}_{m-1}(\mathbf{z})]/2$ is usually used for both $\mathbf{A}_m(\mathbf{z})$ and $\mathbf{A}_{m-1}(\mathbf{z})$. If we follow the assumptions of Gunnar-Farneback Optical Flow and utilize the derived equations, the temporal contrast can be expressed in terms of the displacement vector and the coefficients of the quadratic polynomial:

$$\Delta I_m(\mathbf{z}) = |-2\mathbf{d}_m(\mathbf{z})^{\mathrm{T}}\mathbf{A}_{m-1}(\mathbf{z})^{\mathrm{T}}\mathbf{z} + \mathbf{d}_m(\mathbf{z})^{\mathrm{T}}\mathbf{A}_{m-1}(\mathbf{z})\mathbf{d}_m(\mathbf{z}) - \mathbf{b}_{m-1}(\mathbf{z})^{\mathrm{T}}\mathbf{d}_m(\mathbf{z})|, \quad (26)$$

which clearly shows the interdependence between optical flow and temporal contrast

ORCID iDs

Yuanyuan Yao https://orcid.org/0000-0002-4307-8137
Axel Stebner https://orcid.org/0000-0003-4161-8344
Tinne Tuytelaars https://orcid.org/0000-0003-3307-9723
Simon Geirnaert https://orcid.org/0000-0002-4120-4232
Alexander Bertrand https://orcid.org/0000-0002-4827-8568

References

[1] Luck S J 2014 *An Introduction to the Event-Related Potential Technique* (MIT Press)

[2] Apicella A, Arpaia P, Frosolone M, Improta G, Moccaldi N and Pollastro A 2022 EEG-based measurement system for monitoring student engagement in learning 4.0 *Sci. Rep.* **12** 5857

[3] Zioga P, Pollick F, Minhua M, Chapman P and Stefanov K 2018 "Enheduanna-a manifesto of falling" live brain-computer cinema performance: performer and audience participation, cognition and emotional engagement using multi-brain BCI Interaction *Front. Neurosci.* **12** 191

[4] Aricó P *et al* 2016 Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment *Front. Hum. Neurosci.* **10** 539

[5] Alarcão S M and Fonseca M J 2019 Emotions recognition using EEG signals: a survey *IEEE Trans. Affect. Comput.* **10** 374–93

[6] Ding N and Simon J Z 2012 Emergence of neural encoding of auditory objects while listening to competing speakers *Proc. Natl Acad. Sci.* **109** 11854–9

[7] Lalor E C and Foxe J J 2010 Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution *Eur. J. Neurosci.* **31** 189–93

[8] Biesmans W, Das N, Francart T and Bertrand A 2017 Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 402–12

[9] Vanthornhout J, Decruy L, Wouters J, Simon J Z and Francart T 2018 Speech intelligibility predicted from neural entrainment of the speech envelope *J. Assoc. Res. Otolaryngol.* **19** 181–91

[10] Di Liberto G M, O'Sullivan J A and Lalor E C 2015 Low-frequency cortical entrainment to speech reflects phoneme-level processing *Curr. Biol.* **25** 2457–65

[11] Broderick M P, Anderson A J and Lalor E C 2019 Semantic context enhances the early auditory encoding of natural speech *J. Neurosci.* **39** 7564–75

[12] De Cheveigné A, Wong D D E, Di Liberto G M, Hjortkjær J, Slaney M and Lalor E 2018 Decoding the auditory brain with canonical component analysis *NeuroImage* **172** 206–16

[13] Puffay C, Accou B, Bollens L, Jalilpour Monesi M J, Vanthornhout J, Van hamme H and Francart T 2023 Relating EEG to continuous speech using deep neural networks: a review *J. Neural Eng.* **20** 041003

[14] Reddy Katthi J, Ganapathy S, Kothinti S and Slaney M 2020 Deep canonical correlation analysis for decoding the auditory brain *2020 42nd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE) pp 3505–8

[15] Jalilpour Monesi M, Accou B, Montoya-Martinez J, Francart T and Van Hamme H 2020 An LSTM based architecture to relate speech stimulus to EEG *ICASSP 2020 - 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 941–5

[16] Geirnaert S, Vandecappelle S, Alickovic E, De Cheveigne A, Lalor E, Meyer B T, Miran S, Francart T and Bertrand A 2021 Electroencephalography-based auditory attention decoding: toward neurosteered hearing devices *IEEE Signal Process. Mag.* **38** 89–102

[17] Dmochowski J P, Sajda P, Dias J and Parra L C 2012 Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement? *Front. Hum. Neurosci.* **6** 112

[18] Zhang J R, Sherwin J, Dmochowski J, Sajda P and Kender J R 2014 Correlating speaker gestures in political debates with audience engagement measured via EEG *Proc. 22nd ACM International Conference on Multimedia* (ACM) pp 387–96

[19] Dmochowski J P, Bezdek M A, Abelson B P, Johnson J S, Schumacher E H and Parra L C 2014 Audience preferences are predicted by temporal reliability of neural processing *Nat. Commun.* **5** 4567

[20] Trier Poulsen A T, Kamronn S, Dmochowski J, Parra L C and Kai Hansen L K 2017 EEG in the classroom: synchronised neural recordings during video presentation *Sci. Rep.* **7** 43916

[21] Dmochowski J P, Ki J J, DeGuzman P, Sajda P and Parra L C 2018 Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity *NeuroImage* **180** 134–46

[22] Ki J J, Parra L C and Dmochowski J P 2020 Visually evoked responses are enhanced when engaging in a video game *Eur. J. Neurosci.* **52** 4695–708

[23] Gramfort A *et al* 2013 MEG and EEG data analysis with MNE-Python *Front. Neurosci.* **7** 1–13

[24] Farnebäck G 2003 Two-frame motion estimation based on polynomial expansion *Image Analysis* (*Lecture Notes in Computer Science* vol 2749) eds G Goos, J Hartmanis, J Van Leeuwen, J Bigun and T Gustavsson (Springer) pp 363–70

[25] Bradski G 2000 The OpenCV Library *Dr. Dobb's j. softw. tools prof. program.* **25** 120–3

[26] He K, Gkioxari G, Dollar P and Girshick R 2020 Mask R-CNN *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 386–97

[27] Hotelling H 1992 Relations between two sets of variates *Breakthroughs in Statistics: Methodology and Distribution* (Springer) pp 162–90

[28] Bayro Corrochano E, De Bie T, Cristianini N and Rosipal R 2005 Eigenproblems in pattern recognition *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing and Robotics* (Springer) pp 129–67

[29] Douglas Carroll J 1968 Generalization of canonical correlation analysis to three of more sets of variables *Proc. 76th Annual Convention of the American Psychological Association* pp 227–8

[30] Geirnaert S, Francart T and Bertrand A 2023 Stimulus-informed generalized canonical correlation analysis of stimulus-following brain responses *2023 31st European Signal Processing Conf. (EUSIPCO)* (https://doi.org/10.23919/EUSIPCO58844.2023.10290073)

[31] Hovine C and Bertrand A 2022 MAXVAR-based distributed correlation estimation in a wireless sensor network *IEEE Trans. Signal Process.* **70** 5533–48

[32] De Cheveigné A, Slaney M, Fuglsang S A and Hjortkjaer J 2021 Auditory stimulus-response modeling with a match-mismatch task *J. Neural Eng.* **18** 046040

[33] Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B and Bießmann F 2014 On the interpretation of weight vectors of linear models in multivariate neuroimaging *NeuroImage* **87** 96–110

[34] Virtanen P *et al* (SciPy 1.0 Contributors) 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72

[35] Castellano B 2023 Pyscenedetect (available at: www.scenedetect.com/)

[36] Herbec A, Kauppi J-P, Jola C, Tohka J and Pollick F E 2015 Differences in fMRI intersubject correlation while viewing unedited and edited videos of dance performance *Cortex* **71** 341–8

[37] Nentwich M, Leszczynski M, Russ B E, Hirsch L, Markowitz N, Sapru K, Schroeder C E, Mehta A, Bickel S and Parra L C 2023 Semantic novelty modulates neural responses to visual change across the human brain *Nat. Commun.* **14** 2910

[38] Grosbras M-H, Beaton S and Eickhoff S B 2012 Brain regions involved in human movement perception: a quantitative voxel-based meta-analysis *Human Brain Mapp.* **33** 431–54

[39] Vaessen M J, Abassi E, Mancini M, Camurri A and De Gelder B 2019 Computational feature analysis of body movements reveals hierarchical brain organization *Cereb. Cortex* **29** 3551–60

[40] Rossi A F and Paradiso M A 1999 Neural correlates of perceived brightness in the retina, lateral geniculate nucleus and striate cortex *J. Neurosci.* **19** 6145–56

[41] Hasson U, Nir Y, Levy I, Fuhrmann G and Malach R 2004 Intersubject synchronization of cortical activity during natural vision *Science* **303** 1634–40

[42] Yao Y, Stebner A, Tuytelaars T, Geirnaert S and Bertrand A 2024 Video-EEG Encoding-Decoding Dataset KU Leuven Zenodo (https://doi.org/10.5281/zenodo.10512414)