Pattern Recognition 45 (2012) 2109–2122

Contents lists available at ScienceDirect

Pattern Recognition

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

Decoding stimulus-related information from single-trial EEG responses based on voltage topographies

Athina Tzovara [a,b], Micah M. Murray [a,b,c], Gijs Plomp [d], Michael H. Herzog [d], Christoph M. Michel [e], Marzia De Lucia [a,b,*]

[a] Electroencephalography Brain Mapping Core, Center for Biomedical Imaging of Lausanne and Geneva, Switzerland
[b] Radiology Department, Vaudois University Hospital Center and University of Lausanne, Switzerland
[c] Department of Clinical Neurosciences, Vaudois University Hospital Center and University of Lausanne, Switzerland
[d] Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[e] Functional Brain Mapping Laboratory, Department of Fundamental and Clinical Neuroscience, University Hospital and University Medical School, Geneva, Switzerland

ARTICLE INFO

Available online 20 April 2011

Keywords:
Decoding
Single-trial
EEG
Topographic analysis
Classification
Gaussian Mixture Model

ABSTRACT

Neuroimaging studies typically compare experimental conditions using average brain responses, thereby overlooking the stimulus-related information conveyed by distributed spatio-temporal patterns of single-trial responses. Here, we take advantage of this rich information at a single-trial level to decode stimulus-related signals in two event-related potential (ERP) studies. Our method models the statistical distribution of the voltage topographies with a Gaussian Mixture Model (GMM), which reduces the dataset to a number of representative voltage topographies. The degree of presence of these topographies across trials at specific latencies is then used to classify experimental conditions. We tested the algorithm using a cross-validation procedure in two independent EEG datasets. In the first ERP study, we classified left- versus right-hemifield checkerboard stimuli for upper and lower visual hemifields. In a second ERP study, when functional differences cannot be assumed, we classified initial versus repeated presentations of visual objects. With minimal *a priori* information, the GMM model provides neurophysiologically interpretable features – vis à vis voltage topographies – as well as dynamic information about brain function. This method can in principle be applied to any ERP dataset testing the functional relevance of specific time periods for stimulus processing, the predictability of subject's behavior and cognitive states, and the discrimination between healthy and clinical populations.

1. Introduction

In electroencephalography (EEG) research, averages of the peri-stimulus EEG across trials are typically used to derive ERPs at each recorded electrode and to study evoked neural responses to external stimuli. Averaging across trials improves the signal-to-noise ratio (SNR) and reduces the influence of physiological and instrumental noise because it preserves only those EEG signals that are time-locked to stimulus onset. Recently, it has become increasingly clear that the EEG contains functionally important signals that are not time-locked to stimulus onset and can only be studied at the single-trial level [1–4]. The impact of such unlocked signals is strongest when high-level cognitive factors play a role in task completion [5,6] and when effects of learning and plasticity cannot be excluded [4,7,8]. Importantly, detection of stimulus-related activity at the single-trial level can be used for developing classification strategies to predict the subject's functional state (i.e. learning stage) or his behavioral output [9]. This approach has applications ranging from anticipating subject's preferences in a decision-making task ([10,11] using EEG and [12] using functional magnetic resonance imaging—fMRI) or uncovering hidden intentions in a lie-detection paradigm (see [13] using fMRI).

Whereas decoding approaches are well established in fMRI research [14–17], in the EEG domain they are mainly used in Brain–Computer Interface (BCI) applications and are far less common in ERP studies. This is surprising because the EEG signal, with its high temporal resolution, may contain a great amount of subtle stimulus-related information. By applying multivariate techniques to the single-trial ERPs, we expect to uncover effects that cannot be detected at the average ERP level. Previous attempts to decode stimulus-related signals in single-trial ERPs were based largely on a trial-and-error strategy for selecting

* Correspondence to: Electroencephalography Brain Mapping Core, Lemanic Center for Biomedical Imaging, Centre Hospitalier Universitaire Vaudois, BH 07 081 1, Department of Biology and Medicine, University of Lausanne, 1011 Lausanne, Switzerland. Tel.: +41 21 314 68 28; fax: +41 21 314 46 45.
E-mail address: marzia.de-lucia@chuv.ch (M. De Lucia).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.04.007

those temporal periods that were most discriminant and fixed *a priori* the time-window(s) of interest [10,11].

Here, we propose a general approach for decoding stimulus categories at the single-trial level that requires minimal *a priori* assumptions. Specifically our method estimates the onset and temporal duration of the effect of interest by a data-driven approach, does not need any electrodes selection and is independent of the EEG reference. Therefore, we take full advantage of the voltage measured across electrodes over the scalp, showing that the scalp voltage topography of the EEG contains stimulus-related information that can be detected at the single-trial level and used to decode stimuli. This approach is rooted in a long tradition of topographic analysis of ongoing EEG and ERPs. It has been shown that the brain's electric field at the scalp does not fluctuate randomly across time, but instead remains in stable configurations for periods of several tens of milliseconds, independently of changes in response strength [18–24]. The advantage of using voltage topographies for decoding is their neurophysiological interpretability; changes in the scalp voltage topography forcibly follow from changes in the underlying configuration of intracranial generators [18,20,21].

We previously developed topographic clustering procedures based on a GMM model of single-trial ERP data [25,26]. However, these previous results applying a GMM model were based either on a qualitative analysis of a dataset collapsed across subjects [25] or on one subject's data [26]. Here, we extend our method to identify those stimulus-related topographic features that optimally separate stimulus categories and can therefore be used to classify new sets of single-trials as belonging to one condition or the other. This multivariate decoding strategy identifies which critical voltage topographies (and at which latencies) underlie the functional differences between two experimental conditions.

For validation purposes, in this study we apply our methods on two visual evoked potential (VEP) datasets with well-established voltage topographies that occur at known latencies at the average ERP level [27–29]. Because in these two datasets trial-to-trial variability is expected to be minimal, we can compare the features extracted at the single-trial level with those estimated at the average ERP level and then evaluate the extent to which time-unlocked activity is present at the level of single-trial EEG. In addition, we compare the topographic classification performances based on single-trial and average ERP level and demonstrate that single-trial topographic classification produces the higher classification accuracy. This test is essential to show the added value of single-trial analysis over average ERP and an important step before the classification algorithm can be used in more general contexts.

2. Proposed method for single-trial classification

The classification algorithm described in the following can be applied to any ERP study involving at least two experimental conditions. In its present form, it can be applied to single subjects, each subject being analyzed independently. In what follows we will always refer to an ERP dataset of one subject including two experimental conditions. This dataset consists of the concatenation of peri-stimulus EEG epochs (*trials*). Each trial can be of any length, but usually the post-stimulus period is much longer that the pre-stimulus period. Moreover, at each time point, the data across the electrode montage constitute a vector of voltage measurements $\boldsymbol{m}=\{m_1,m_2,\ldots,m_N\}$, where N is the total number of electrodes (Fig. 1a). We will refer to this vector as a topography. These datasets are pre-processed by normalizing each topography by its instantaneous Global Field Power (GFP) [30,31]. This normalization makes the classification algorithm solely dependent on the overall shape of the topography irrespective of its instantaneous strength [19,20]. In the following, $\boldsymbol{m}$ will always represent the GFP-normalized generic topography at one time point. The algorithm is comprised of two main steps: the training and the test phase. The training dataset is the part of the data that is used for estimating the ERP model. In the following 'T' indicates the total number of trials in the training dataset, and without loss of generality we assume that the two datasets have the same number of trials. The test dataset, including a selection of trials that did not overlap with those of the training dataset, is used for selecting the optimal model parameters. This split in training and testing is done for each of the two conditions, separately. Moreover, a set of trials (validation dataset) is kept completely independent of the training and testing datasets in order to objectively measure our methods' performance in real-life decoding applications.

2.1. Training phase

2.1.1. Gaussian Mixture Model estimation

The first step of our analysis comprises a modeling of the ensemble of topographies in the training dataset for each experimental condition. At this stage of the analysis, all the available topographies are pooled together disregarding the latencies at which they are observed and the trial to which they belong (Fig. 1a). Therefore, their temporal order is not relevant at this point.

To reduce our ensemble of topographies to a number of representative *template maps*, we propose a GMM probability distribution in an N-dimensional space (Fig. 1b)

$$P(\boldsymbol{m}|\boldsymbol{\mu},\boldsymbol{\sigma})=\sum_{k=1}^{Q} p_k P_k(\boldsymbol{m}|\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k) \tag{1}$$

where P_k is the kth Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\sigma}_k$, p_k is the prior probability of the class label k and Q is the total number of Gaussians. In the following, we will refer to the means – $\boldsymbol{\mu}_k$ – of the Gaussians as template maps, and for simplicity we will replace the notation "$\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k$" with "c_k" to indicate the kth Gaussian within the GMM.

In order to estimate the GMM distribution, we use an expectation-maximization procedure [32,33] that iterates the estimation of the model's parameters (priors, means and covariance matrices) in order to minimize the error function, or equivalently maximizing the likelihood $\mathcal{L}$

$$E=-\ln\mathcal{L}=-\sum_{n}\ln\left\{\sum_{k=1}^{Q} P(\boldsymbol{m}|c_k)p_k\right\} \tag{2}$$

where the index n spans the total number of topographies in the training dataset, i.e. the total number of topographies in one trial multiplied by the total number of trials in the training dataset for one condition.

In order to initiate this algorithm, a first guess of the means and covariance matrices needs to be provided. We consider as initial means those obtained by a k-means clustering algorithm [33]. The initial estimation of covariance matrices is obtained by considering the topographies belonging to each cluster as estimated by the k-means algorithm. The initial values for the priors, p_k, are obtained by the relative number of topographies for each cluster. Due to the limited number of training samples, we reduce the number of free parameters by constraining the covariance matrix to be diagonal. Although in principle this assumption implies independency among different electrodes, the use of a mixture model can still account for dependency at a global level [34,35].

The above mentioned procedure of computing a GMM model is performed for each condition separately, so we end up with one model per condition. The only *a priori* hypothesis is the total number of Gaussians, Q, in the model (in the example of

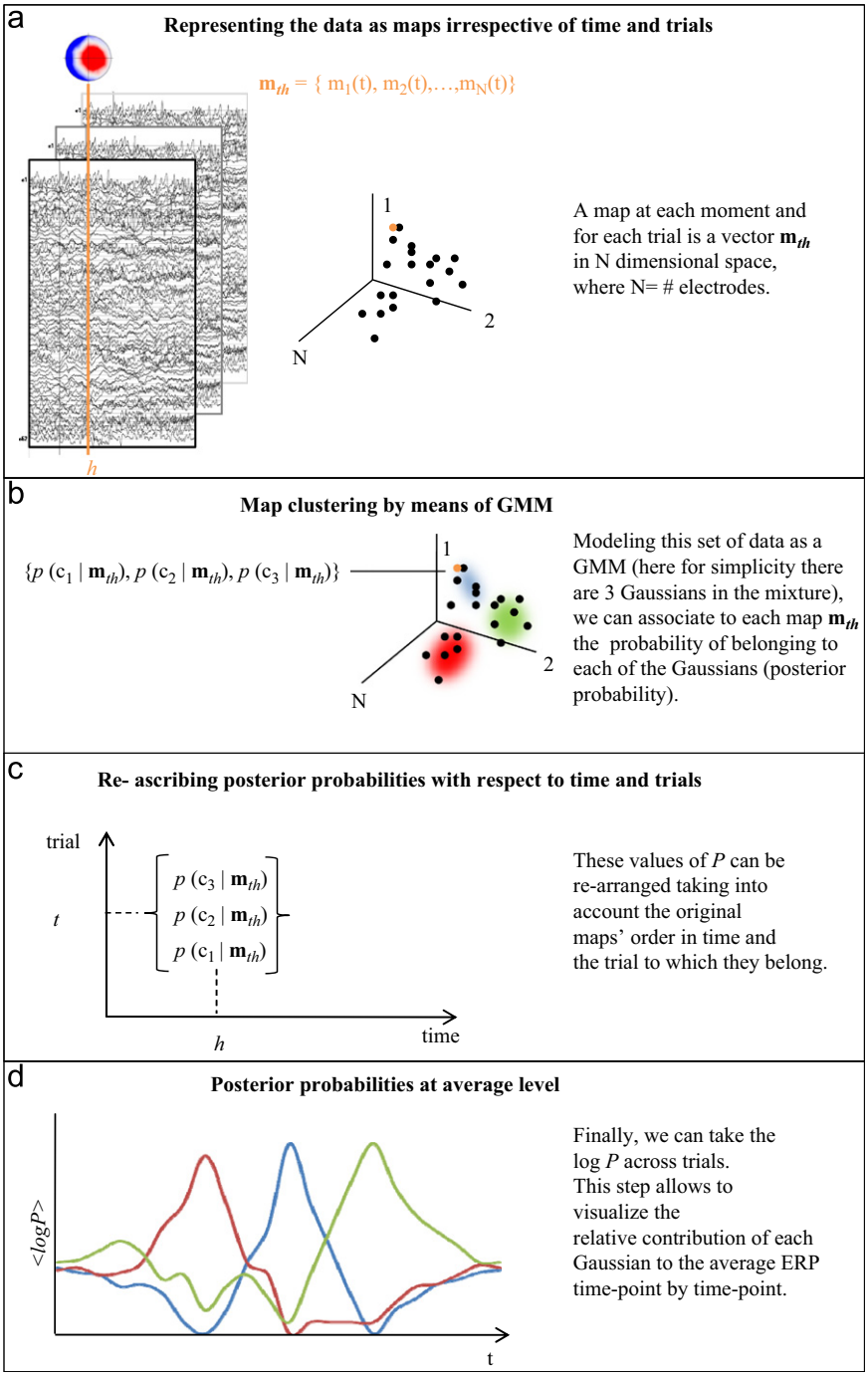

Fig. 1. Schematic representation of how we model the ERP dataset of each experimental condition by means of GMM. (a) At each time-point, the topography is represented as a vector in N-dimensional space (where N=#electrodes) without taking into account the latency h and the trial to which it belongs. (b) Example of a GMM with three Gaussians in the mixture. After evaluating the model's parameters, each topography is assigned a set of posterior probabilities, the number of which equals that of the Gaussians in the mixture. These probabilities quantify the confidence with which each topography is assigned to each Gaussian. (c) The set of posterior probability values are re-arranged time-point by time-point and re-assigned to each of the original trials. (d) Computing the posterior probabilities across trials (Eq. (4)), we can quantify the degree of presence of each template map in time across trials (i.e. both within and between experimental conditions).

Fig. 1b, $Q=3$). In the following, we will consider two generic values $Q1$ and $Q2$ for the first and the second condition, respectively. However, we will explain below in a dedicated paragraph how we select the optimal values for $Q1$ and $Q2$ (see *Optimizing the total number of Gaussians*).

2.1.2. Evaluating the posterior probabilities of single-trial ERP data

We use GMM models to assign each topography of the original dataset to one of the Gaussians c_k. For this purpose we computed the posterior probabilities, normally used in classification problems, because misclassification error is minimized by choosing the class having the largest posterior probability [33].

The posterior probabilities for each Gaussian in the mixture are defined as in the following:

$$P(c_k|\boldsymbol{m}) = \frac{P(\boldsymbol{m}|c_k)p_k}{p(\boldsymbol{m})} \tag{3}$$

where $p(\boldsymbol{m})$ is the unconditional density function, i.e. the density function for $\boldsymbol{m}$ irrespective of the Gaussian c_k. In order to

investigate stimulus-related information, we need to rearrange the posterior probabilities to their original temporal order in the data, providing a new representation of the single-trial ERPs in time and across trials. For each trial, this rearrangement generates a time series of the posterior probability for each template map in the mixture (Fig. 1c).

The posterior probabilities across trials[1],

$$\sum_{t=1}^{T} \log(P(c_k|\boldsymbol{m}_{th})) - (T-1)\log(p_k) \quad (4)$$

typically reveal a pattern of presence of a given template map that is structured in time (Fig. 1d, 4a, c and 7a), where $\boldsymbol{m}_{th}$ is the topography of tth trial, at time-latency h. This posterior probability is computed at each latency and for each Gaussian c_k in the GMM model for each experimental condition.

In the following, we will make the assumption that the p_k are all equal. In fact, this posterior probability refers to just one point in time h, whereas the priors p_k were computed independently of time. Note that, given this assumption, the second term in Eq. (4) is constant and equal for both conditions; therefore it will not play a role in the algorithm implementation.

At this point we are able to compare the two conditions by taking advantage of the model estimated on each of the two conditions separately.

2.1.3. Identifying discriminative time periods between conditions

We compute the posterior probabilities of each of the two models and datasets. This is equivalent to considering a joint model for the entire dataset given by the sum of each of the two GMMs with equal weights [36,37]. In the following we indicate with $\boldsymbol{m1}_{th}$ and $\boldsymbol{m2}_{th}$ the topography of Condition 1 and Condition 2, respectively, appearing at the trial t and latency h; c_i and c_j refer to two generic Gaussians in the GMMs for the two conditions (and therefore $1 \le i \le Q1$ and $1 \le j \le Q2$). Based on the two models and datasets, we compute $P(c_i|\boldsymbol{m1}_{th})$ and $P(c_j|\boldsymbol{m2}_{th})$.

At each time-point h, we consider one Gaussian per model, i.e. the one that yields the highest posterior probability across trials (Eq. (4); Fig. 2a, cyan curves):

$$i_h = \underset{i \in 1:Q1}{\operatorname{argmax}} \left\{ \frac{\sum_{t=1}^{T} \log P(c_i|\boldsymbol{m1}_{th})}{T} \right\} \quad (5)$$

$$j_h = \underset{j \in 1:Q2}{\operatorname{argmax}} \left\{ \frac{\sum_{t=1}^{T} \log P(c_j|\boldsymbol{m2}_{th})}{T} \right\} \quad (6)$$

where $i \in 1:Q1$ and $j \in 1:Q2$ denotes the range of maps for the first and second model, respectively. If the two GMMs capture the difference between the two conditions, we expect that the Gaussians c_{i_h} and c_{j_h} are characterized by different parameters (i.e. different template maps), that is to say c_{i_h} and c_{j_h} are very specific to Condition 1 and Condition 2, respectively. However this might be the case only along a specific time period within the overall trial. Therefore, we refine our strategy by looking only at those time points at which c_{i_h} and c_{j_h} are selectively representative of Condition 1 and Condition 2. It is worth noting that the parameters of the Gaussians c_{i_h} and c_{j_h} are estimated on the overall datasets and therefore do not depend on the specific latency 'h'; however, we use the indexes i_h and j_h to emphasize that the selection of the Gaussians depends on h.

We use the Bayes factors (BF; Eqs. (7) and (8)) to identify periods over which each condition is better explained by its own

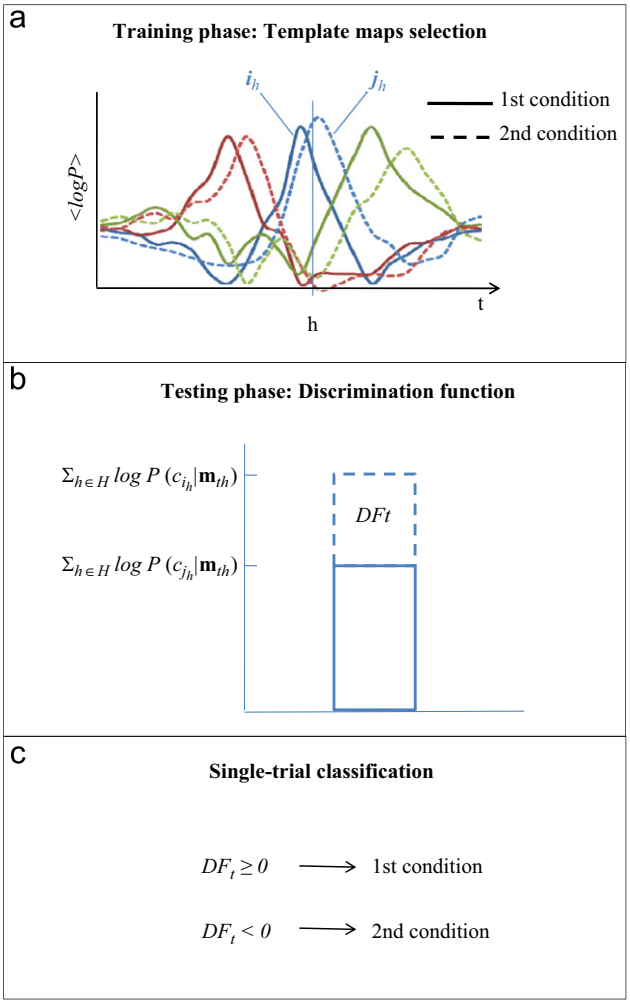

Fig. 2. Schematic representation of how we select the template maps that are used for building the discrimination function and performing classification. (a) The posterior probabilities of the training datasets are computed based on the corresponding model (continuous and dashed lines). At each latency 'h' the template map yielding the highest posterior probability across trials is selected (i_h and j_h in the Eqs. (5) and (6)). (b) For each trial of the testing dataset, we compute the logarithm of posterior probability for the map i_h and the posterior probability for map j_h, averaged over H time-points responsible for the difference between conditions (see Section 2.1.3). The discrimination function is defined as the difference of these two values. (c) The trial is classified as belonging to the first condition if the discrimination function is equal to or greater that zero, else it is classified as belonging to the second.

model, across trials:

$$BF1 = \frac{P(c_{i_h}|\boldsymbol{m1}_{th})}{P(c_{j_h}|\boldsymbol{m1}_{th})} \quad (7)$$

$$BF2 = \frac{P(c_{j_h}|\boldsymbol{m2}_{th})}{P(c_{i_h}|\boldsymbol{m2}_{th})} \quad (8)$$

Because we have considered equal priors, the ratio of the likelihoods in Eqs. (7) and (8) is the same as the ratio of the posteriors probabilities. The BF can be computed at each latency and for every trial, separately, and gives the confidence with which we can assign a specific observation to a given template map of one of the models. This quantity in Bayesian statistics is similar to the p-value in classical statistics [38]. In fact, if the BF is greater than one it means that the template map that appears in the numerator (i.e. c_{i_h} for BF1 and c_{j_h} for BF2) better explains the specific observation ($\boldsymbol{m1}_{th}$ for BF1 and $\boldsymbol{m2}_{th}$ for BF2). Moreover, if

[1] Note that without loss of generality in the following we use the logarithm of the posterior probabilities expressed in Eq. (3) and the equivalence: $\log(\prod_{t=1}^{T} P(c_k|\boldsymbol{m}_{th})) = \sum_{t=1}^{T} \log(P(c_k|\boldsymbol{m}_{th}))$ and we will use the notation $\langle \log P \rangle$ in the figures to indicate this sum.

BF1 is found to be greater than 20, this can be interpreted as a strong evidence of c_{i_h} over c_{j_h} [39].

Here, we computed the BF1 and BF2 at every trial and time latency and retained a specific latency if BF1 or BF2 was greater than 20 in at least 60% of the trials. This 60% threshold was chosen so as to obtain above chance levels results when it comes to classification of new trials. Overall, we obtain H time points over which the two conditions differ.

2.2. Testing phase

2.2.1. Discrimination function

The posterior probabilities defined in the previous paragraph, $P(c_{i_h}|\boldsymbol{m}_{th})$ and $P(c_{j_h}|\boldsymbol{m}_{th})$, are computed on the test dataset and along the temporal period H that was estimated in the training phase. Based on the Bayes factor, the discrimination function for the trial t (equivalent to a log-likelihood ratio) is defined as: (Fig. 2b, c)

$$DF_t = \sum_{h \in H} [\log P(c_{i_h}|\boldsymbol{m}_{th}) - \log P(c_{j_h}|\boldsymbol{m}_{th})] \quad (9)$$

The idea is that we assign a trial from the test dataset to Condition 1 if the posterior probability of the most likely Gaussian for the first model (Eq. (5)) is higher than the posterior probability of the most likely Gaussian for the second model (Eq. (6))[2], and vice versa.

The discrimination function is restricted to those time-points where the two conditions were most different so as to enhance their separation. This is the reason why we only use the H time-points obtained from the abovementioned comparison instead of the overall trial.

2.2.2. Optimizing the total number of Gaussians

In order to find the optimum values for $Q1$ and $Q2$, a set of models is generated with $Q1$ and $Q2$ number of Gaussians, respectively, with each of these two parameters ranging from three to eleven for both experiments. The optimal values of $Q1$ and $Q2$ are decided on the basis of mean classification performance across ten splits in the cross-validation procedure. Classification performance is computed for each test dataset in the cross-validation by manipulating the threshold in the discrimination function (Eq. (9)) and evaluating the area under the ROC curve (AUC; [40]).

Because we have considered ten splits of the data, for each pair of models with $Q1$ and $Q2$ number of Gaussians, we obtain ten values of AUC. The best number of Gaussians is determined so as to maximize the mean AUC across the ten splits. This procedure is done for individual trials or sub-sets of trials when the performance obtained when classifying single-trials is not significantly above chance (in this case the notation 't' specifies the average of the sub-sets).

In the following for completeness we will also report the average absolute accuracy computed across the ten data splits although it is not used for parameters' selection. This absolute accuracy is computed by setting a threshold in the discrimination function for each split separately and evaluating the percentage of correctly classified trials in the test dataset.

2.2.3. Classification accuracy on the validation dataset

So far, we have used a cross validation procedure for selecting an optimal model. The corresponding classification performance is only indicative of the classifier's performance because it is obtained on the same dataset used for optimizing models' parameters. A more reliable assessment of the classification performance is based on a validation dataset that is not used either for training or testing. This classification performance is quantified with the AUC on a set of completely unknown trials.

2.3. Single-trial classification based on the average ERP model

In order to deeper investigate the classification results and the actual advantage we gain by using the single-trial information, we considered the following approach: instead of using all the single-trial data of the training dataset to compute the GMM models, we used the ERP averaged data (of the training dataset). Apart from the different model estimation, we repeat the same procedure as described above (see Section 2). The only difference is that we cannot take advantage of the trial repetitions, so we consider as periods of difference H the periods where the Bayes factor for one or the other condition's average ERP is higher than 20. In the testing phase, we classify single-trials as described in the previous section.

3. Experimental paradigm

The data analyzed in this study were acquired from two separate VEP experiments. The first entailed passive presentation of checkerboard stimuli to each of the four visual quadrants (hereafter, "Checkerboard Experiment"). The second entailed active discrimination of novel versus repeated line drawings of common objects (hereafter, "Priming Experiment"). Full details for the paradigm and data acquisition can be found in published studies of the VEPs [29,27,28].

3.1. Checkerboard Experiment: subjects, stimuli and task

We used data from 4 subjects (1 female), aged 18–28 years with normal or corrected-to-normal vision. The subjects were all right-handed [41]. Informed consent was obtained before the experiment and the procedures were in accordance with the Declaration of Helsinki and approved by the local ethics committee.

Stimuli were presented on a ViewSonic G90f+CRT screen with a 75 Hz refresh rate. Maximal screen luminance was $\sim$80 cd/m^2. Subjects were seated one meter from the monitor in an electrically shielded room. The checkerboard stimuli consisted of quarter annuli with an inner radius of 1° of visual angle and an outer radius of 8° from the center of the screen. The annuli edges were offset 3° from the horizontal and vertical meridian to limit contamination across quadrants in the evoked response. The annuli were filled with a $11^\circ \times 8^\circ$ rectangular checkerboard in which the size of each rectangle was proportional to the annulus width.

Subjects were instructed to fixate a green fixation point (5.5 arcmin) at the center of the screen and to press a hand-held button when its color changed to red. This occurred approximately 10 times per recording run at random intervals. We instructed observers to emphasize accuracy over speed and to refrain from moving and blinking during the recording. Checkerboard stimuli appeared for 147 ms in one quadrant of the visual field at a time with a randomized inter-stimulus interval (ISI) of 300–800 ms. Presentation order was randomized with the restriction that the same quadrant was never stimulated twice in a row. Two blocks of four runs of 240 trials each were recorded for each subject. In total, 480 checkerboards were presented to each quadrant.

[2] In practice, each term in the sum of Eq. (9) is computed as an average along a time window $u \ll H$ to allow for a possible jitter in time of the appearance of the template maps. In the following we will set u equal to 7 ms.

3.2. Checkerboard Experiment: EEG acquisition

Continuous EEG data were recorded with a BioSemi Active Two system (BioSemi, Amsterdam, the Netherlands) using 192 Ag–AgCl sintered active electrodes. The recording was referenced to the CMS–DRL ground, a feedback loop that keeps the montage potential close to amplifier zero. The electro-oculogram (EOG) was recorded with electrodes 1 cm above and below the right eye, and 1 cm lateral to the left and right outer canthus. The recording sampling rate was 2048 Hz (offline the data were down-sampled to 512 Hz). There were 200 trials in response to each visual quadrant that were included in the analyses. Peri-stimulus EEG epochs spanned from 98 ms pre-stimulus to 293 ms post-stimulus. The remainder of the pre-processing steps was common to both experiments and is described below.

3.3. Priming Experiment: subjects, stimuli and task

Four paid volunteers (1 female), aged 23–27 years provided written, informed consent to participate in the experiment, the procedures of which were approved by the Ethics Committee of the University Hospital of Geneva. All subjects were right-handed [41], had no neurological or psychiatric illnesses, had normal or corrected-to-normal vision and reported normal hearing. Subjects performed a continuous recognition task comprised of equal numbers of initial and repeated presentations of line drawings (cf. Fig. 1 in [27]). This task had subjects indicate whether each visual stimulus was appearing for the first time or had appeared previously. Visual stimuli were comprised of line drawings of common objects selected from either a standardized set [42] or obtained from an online library (http://dgl.microsoft.com) and modified to stylistically resemble those from the standardized set. Images appeared black on a white background and were centrally presented for 500 ms on a computer monitor (Sony Trinitron Multiscan model no. GDM-20SE1VT) located 150 cm from the subject. While the original experiment included four conditions to examine the impact of multisensory stimulus encoding on visual memory retrieval, here we considered only those trials where visual stimuli were never coupled with sounds on their initial or repeated presentation. Stimuli were blocked into a series of 136 trials, with equal likelihood of initial and repeated presentations. During a block of trials, each image was repeated once, independently of how the image was initially presented. The average number of trials between the initial and repeated presentation of any given stimulus was 13 (± 3) images. The timing of trials was such that stimuli were presented for 500 ms, followed by a 1200–1500-ms period of randomized stimulus onset asynchrony (SOA). Each subject completed eight blocks of trials.

3.4. Priming Experiment: EEG acquisition

Continuous EEG was acquired with a Geodesics Netamps system (Electrical Geodesics, Inc., USA) from 123 scalp electrodes (impedances $< 50\ k\Omega$; vertex reference; 500 Hz digitization; band-pass filtered 0.1–200 Hz). After pre-processing 190 trials from each experimental condition and subject were included in the analyses. Peri-stimulus EEG epochs spanned 100 ms pre-stimulus to 500 ms post-stimulus onset.

3.5. EEG pre-processing common to both experiments

Pre-processing of the EEG data was performed using the Cartool software (http://brainmapping.unige.ch/Cartool.htm). Data were band-pass filtered 0.1–40 Hz using a Butterworth filter with -12 db/octave roll-off. A semi-automated $\pm 100\ \mu V$ artifact rejection criterion was applied to all scalp channels. Data were also visually inspected to identify electrodes with poor electrode-skin resistance or other noise transients. Trials with blinks or containing eye movements were excluded. Data were re-referenced offline to the common average reference. For the Checkerboard Experiment there were 200 trials accepted per visual quadrant (upper left, upper right, lower left and lower right) and subject. For the Priming Experiment there were 190 trials accepted per condition (initial and repeated) and subject. Data were then interpolated using 3-D splines [43] and then down-sampled to a 62-channel montage that is comprised of the 10–10 electrode positions. Note that no pre-stimulus baseline correction was applied. By contrast, each topography was normalized to its instantaneous global field power [30,31] so as to remove response strength modulations.

For each experiment, the data were separated into training and test datasets. The training dataset included 90% of the trials for each condition and subject (i.e. 180 for the Checkerboard Experiment and 171 for the Priming Experiment). The test dataset was comprised of the remaining trials. This separation between training and test datasets was performed 10 times in a way that the test datasets were all independent (i.e. did not contain any overlapping trials amongst each other). This was done so as to cross-validate the results as detailed below. In addition, this procedure is in agreement with guidelines to avoid circularity in statistical analyses (e.g. [44]). In the following we will use the terms split or shuffle to indicate each subset of the data that was used for one model training.

4. Results

4.1. Checkerboard Experiment

4.1.1. Average VEP

Visual inspection of the grand-average VEPs in response to stimuli at each visual quadrant revealed the typical pattern of amplitude modulations corresponding to classical VEP components [45–47]. Specifically, electrode Pz exhibited an early peak at 64 ms post-stimulus onset for all four conditions (Fig. 3a and b). This component, often referred to in the literature as the C1 component, was positive for the lower visual field and negative for the upper. The subsequent P1 and N1 components are shown at exemplar electrodes P7 and P8 together with the corresponding topographies on the scalp. The P1 component had a positive focus over the contralateral scalp, here peaking at $\sim$90 ms. The N1 component had a bilateral posterior negative distribution, with a peak at $\sim$160 ms post-stimulus onset.

4.1.2. GMM model estimation and difference between experimental conditions

Parameter selection for the GMM was cross-validated for each subject for pairs of experimental conditions. Here, contrasts always entailed left versus right visual field presentations for upper and lower quadrants, separately. The selected total number of Gaussians in the mixture was in the range of three to nine for each of the two conditions in a given contrast.

Visual inspection of the resulting posterior probabilities across trials revealed a highly structured modulation of template maps appearance during the post-stimulus period for all conditions and subjects (see Fig. 4 for the results from an exemplar subject and Supplemental Fig. 1 for results from the other 3 subjects). Based on these posterior probabilities for each shuffle of the classification test, we found temporal differences starting at $\sim$70 ms post-stimulus onset (i.e. during the C1 component of the VEP described above). For the exemplar subject that is shown in Fig. 4, this difference persisted up to 120 ms, covering also the latency of the

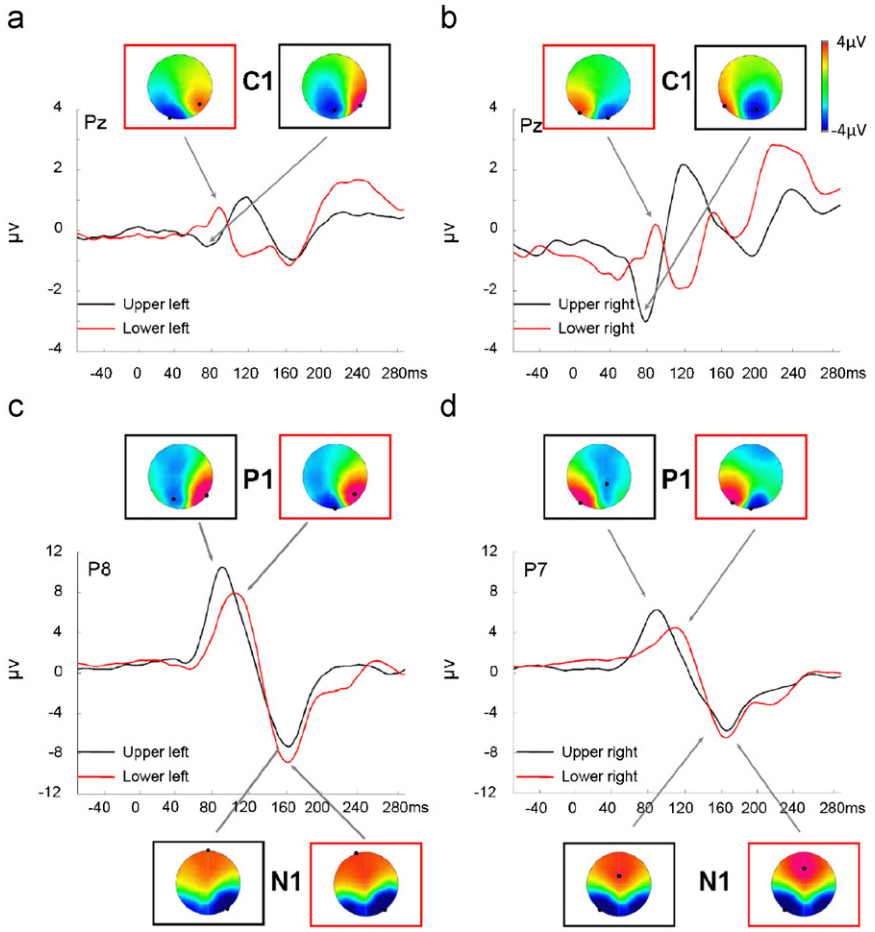

Fig. 3. Grand average VEPs are shown at three exemplar electrodes and for the four conditions corresponding to left (panels a and c) and right (panels b and d) visual field presentations and for upper and lower quadrants (black and red lines, respectively). Voltage topographies are shown at the same latencies of the peaks of the amplitude modulations for each of the electrodes. These peaks correspond to the classical VEP components often referred to as C1 (panels a and b), N1 and P1 (panels c and d). Black dots superimposed on the voltage maps correspond to the maximum and minimum of the voltage values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

P1 component. Also a later difference around 140–180 ms was found during the time interval corresponding to the N1 component. The periods of difference for the other three subjects are shown in Supplemental Fig. 1. The thick yellow line on the *x*-axis of Fig. 4 shows those time intervals for which the difference between conditions was found. Moreover, this difference was observed in at least 8 of the 10 shuffles.

4.1.3. Classification results

The average ROC curves and their range across shuffles are shown in Fig. 5 (mean $\pm$ s.e.m. across shuffles is shown in the inset for each of the ROC curves) and are summarized in Table 1 (first columns). For all four subjects, we found an average AUC significantly higher than chance (unpaired *t*-test, $p < 0.0001$). The average AUC across subjects was 0.80 and 0.81 for upper and lower quadrants, respectively. The absolute accuracy of classification across shuffles was 0.73 on average for both upper and lower quadrants. The classification accuracies were also above chance level for every subject (unpaired *t*-test, $p < 0.0001$). In the validation dataset (including thirty trials per condition and per subject) the average AUC was 0.73 and 0.72 for upper and lower quadrants, respectively.

Moreover, we tested the robustness of these results and the effect of initialization on the computed GMM models for the exemplar subject and the selected total number of Gaussians. The models were recomputed with ten different initializations of the *k*-means algorithm across the ten shuffles of the training dataset [48], and classification was performed on the test datasets, as described above. For the upper visual field the AUC was on average across shuffles and across the ten initializations 0.81 ± 0.003, ranging from 0.80 up to 0.83. For the lower visual field the mean AUC was 0.85 ± 0.003, or ranging from 0.84 to 0.86. In the results we report, this subject had an average across shuffles AUC value of 0.84 and 0.85 for upper and lower visual fields, respectively.

4.1.4. Classification based on the average ERP model

In the case where we re-trained the models using the average ERP data and classified the single-trials of the test datasets (Table 1a and b, second columns), there was a significant drop in classification performance for three out of four subjects, for both upper and lower visual fields (Table 1a and b, third columns; paired *t*-test, $p < 0.05$). More specifically, we remind the reader that for the upper visual field the average AUC when using the models trained on the single-trial data was 0.80. When using the models trained on the average ERPs it was 0.67 on average, resulting in an average drop of 17%. For the lower visual field the average AUC using the single-trial models was 0.81 and when using the average ERP models was 0.73 and had an average drop of 10%.

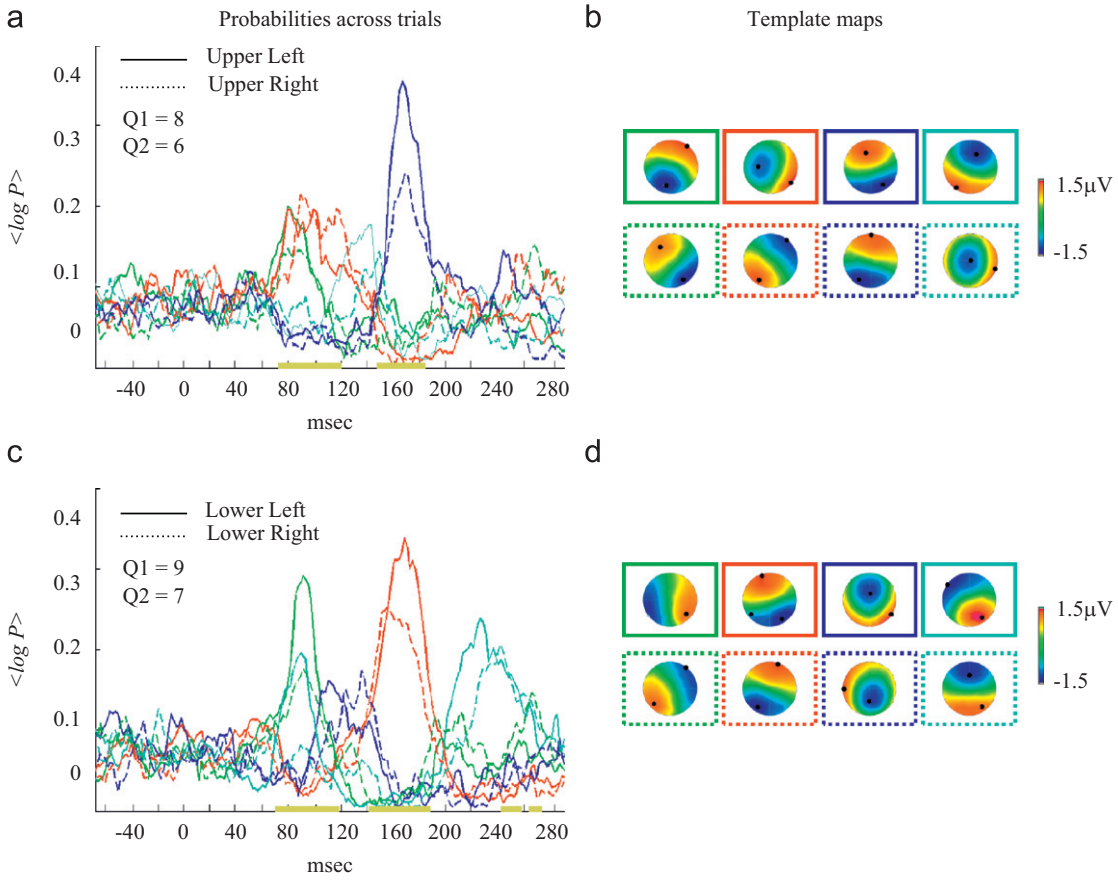

Fig. 4. Posterior probabilities across trials that were used for classification from one exemplar subject in the Checkerboard Experiment. (a) Logarithm of the posterior probabilities of the upper left (continuous line) and upper right (dashed line) visual field presentations and related periods of difference between conditions (thick yellow line on the *x*-axis) (the values of $\langle \log P \rangle$ are normalized so as to range between 0 and 1 (see also footnote 1)); these two sets of probabilities are based on their corresponding GMM. The values of *Q*1 and *Q*2 shown in the inset are the total number of Gaussians that provide the maximum classification accuracy across shuffles for this exemplar subject. The number of posterior probabilities we show are less than the respective values *Q*1 and *Q*2 as we only show those ones that were used for building the classification function. As explained in the text, they correspond to those posterior probabilities that are higher than the others at least in one time-frame where a difference across conditions was found. Color assignment is only for display purposes, it does not carry any information used by the classifier. (b) Template maps, previously normalized by Global Field Power, corresponding to the posterior probabilities that were used for the classification (black dots indicate the maximum and minimum of the voltage values). Frames' colors correspond to the posterior probabilities shown in panel a. Panels c and d display the corresponding results for the lower quadrants, using the same conventions as in panels a and b. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

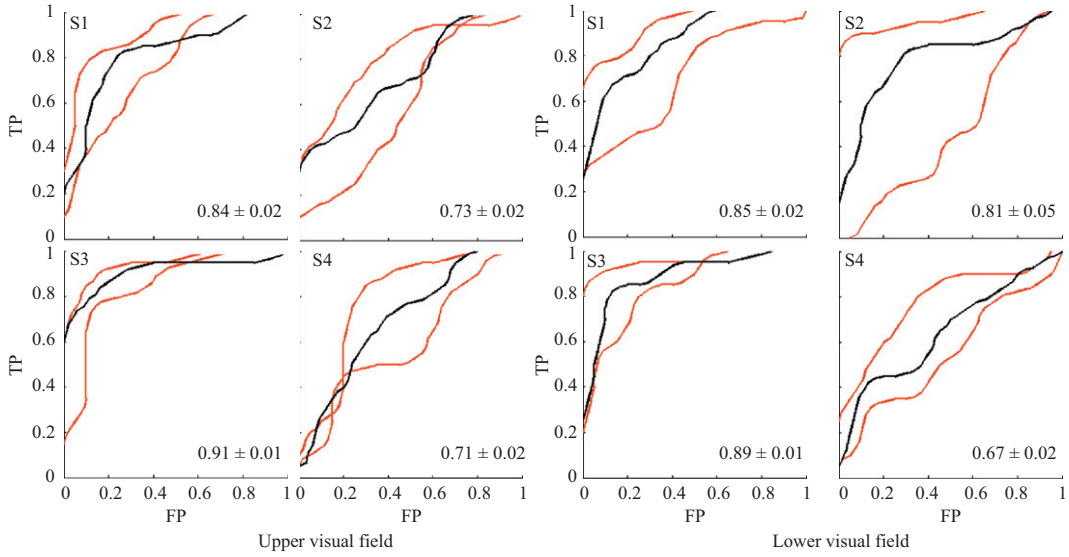

Fig. 5. Average ROC curve for the upper visual field contrast (on the left) and the lower visual field contrast (on the right) are shown in black for each subject (FP and TP stand for rate of false positive and rate of true positive, respectively). In the same panels, pairs of red ROC curves provide the range of the ROC curves across shuffles, representing the worst and the best case. Each of these mean ROC curves was obtained by selecting the GMM models that maximize the average AUC. Mean values across shuffles of the average AUC is shown (mean ± s.e.m.). The exemplar subject shown in Fig. 4 corresponds to S1 for both upper and lower visual field (the others are shown in the Supplemental Material 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1.5. Comparison between template maps and average ERP maps

Posterior probabilities across trials as shown in Fig. 4 provide an estimation of the frequency of each template map at a given latency across trials. For example, a posterior probability across trials of 0.4, indicates that in 40% of the trials the corresponding template map is the one that best fits the data at that latency. More generally, we can reconstruct the topography at a specific time point across trials as a sum of the template maps, each of which is weighted by the corresponding posterior probability. This 'reconstructed ERP' can then be compared to the actual average ERP observed as a function of time. To quantify similarity between these two sets of maps, we computed the spatial correlation at each time point over the post-stimulus period. We computed the average correlation along this post-stimulus period for each subject and condition, separately. The spatial correlations ranged between 0.70 and 0.81 for three out of four subjects and from 0.60 to 0.73 for the fourth (see Supplemental Material 4 for an overview of these maps). We also computed the Global Explained Variance (GEV; [21]) in the same post-stimulus period for each condition. In three of the subjects, values ranged from 0.58 to 0.84 (and from 0.43 to 0.57 in the fourth subject). It should be noted that these values were computed using the reconstructed average from the model whose total number of template maps were selected as explained above; the models were not optimized so as to best fit each of the two conditions; a point to which we return in the 'Discussion Section'.

Table 1
AUC values for all the Checkerboard Experiment based on the single-trial models and the average ERP models. The first column displays the AUC (± standard error) in the case where the classification is based on models computed using all the single-trial data, while the second shows the results using the average ERPs. The relative change column refers to the relative difference in the AUC, when computed using the average ERP model in comparison with the case where we use the single-trial models (asterisks indicates those differences that were significant; paired t-test, $p < 0.05$).

	Single-trial	Average ERP	Relative change (%)
(a) *Checkerboard Experiment—upper visual field*			
S1	0.84 ± 0.02	0.68 ± 0.03	−21*
S2	0.73 ± 0.02	0.63 ± 0.02	−15*
S3	0.91 ± 0.01	0.68 ± 0.04	−29*
S4	0.71 ± 0.02	0.69 ± 0.04	−3
(b) *Checkerboard Experiment—lower visual field*			
S1	0.85 ± 0.02	0.78 ± 0.03	−9*
S2	0.81 ± 0.05	0.72 ± 0.03	−12*
S3	0.89 ± 0.01	0.77 ± 0.05	−15*
S4	0.67 ± 0.02	0.63 ± 0.03	−6

4.2. Priming Experiment

4.2.1. Average VEP

Visual inspection of the grand-average VEPs in response to initial and repeated stimulus presentations revealed the typical pattern of amplitude modulations corresponding to classical P1 and N1 VEP components, though no clear C1 component was observed (Fig. 6). Electrode Pz exhibited P1 and N1 peaks at ~115 and ~170 ms, respectively. The topographies of these components were highly similar across conditions. From ~250 ms post-stimulus onset onwards, there was a clear amplitude modulation between conditions at this electrode, which is in accord with published examples of repetition suppression (e.g. [49] for review). The topography over this time period, although similar, appeared to differ subtly between conditions (in contrast to the more striking topographic differences in the Checkerboard Experiment above).

4.2.2. GMM model estimation and differences between experimental conditions

Parameter selection for the GMM was cross-validated for each subject for pairs of experimental conditions. Here, we contrasted initial versus repeated presentations of identical object stimuli. The optimal number of Gaussians in the mixture was in the range of four to ten for each of the two conditions. The resulting

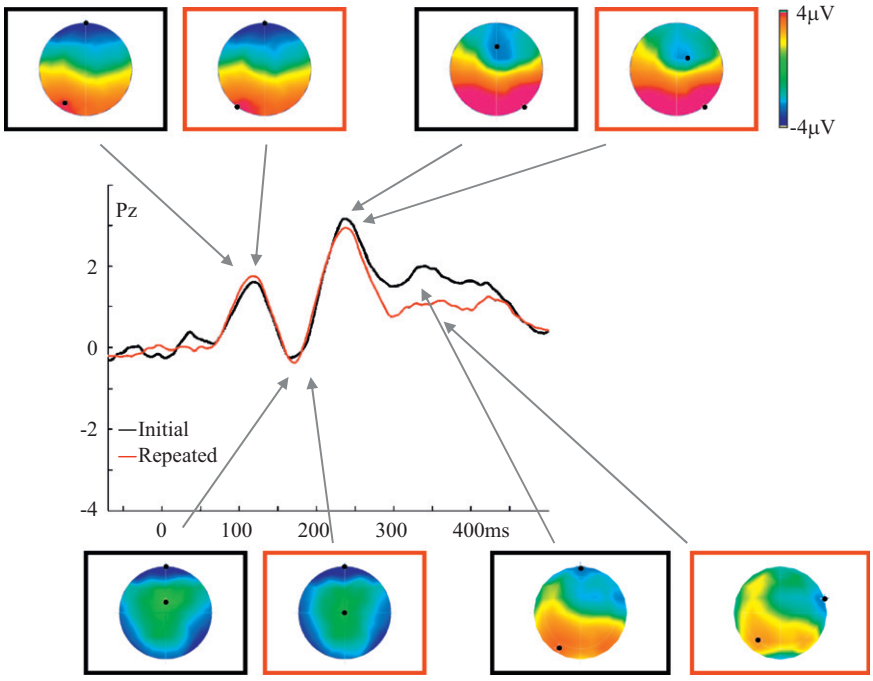

Fig. 6. Grand average VEP at one exemplar electrode and voltage topographies at the main peaks of amplitude modulation are shown for the initial presentation (black line) and repeated presentation (red line) of visual objects in the Priming Experiment. Black dots superimposed on the voltage maps correspond to the maximum and minimum of the voltage values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

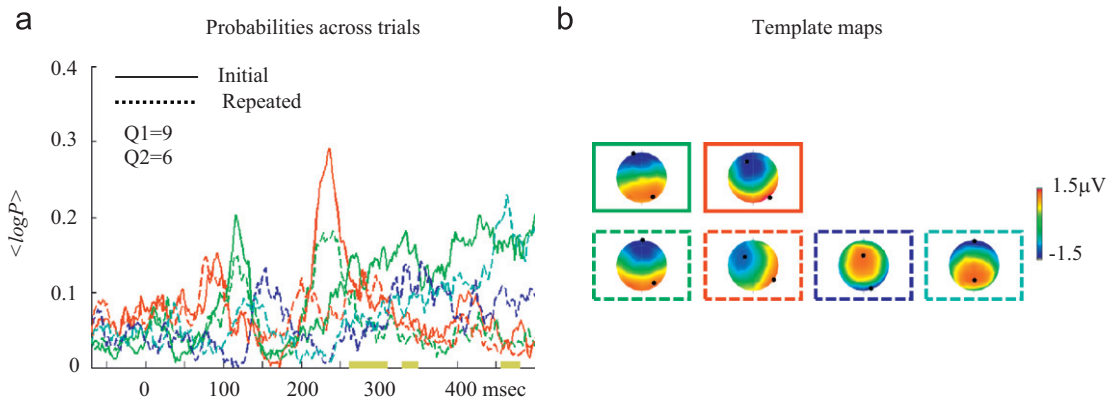

Fig. 7. Posterior probabilities across trials of the Priming Experiment and corresponding template maps. (a) Logarithm of the posterior probabilities of the initial presentation (continuous line) and repeated presentation (dashed line) of identical objects and related periods of difference between conditions (thick yellow line on the *x*-axis) (the values of $\langle \log P \rangle$ are normalized so as to range between 0 and 1 (see also footnote 1)); these two sets of probabilities are based on their corresponding GMM. The values of *Q*1 and *Q*2 shown in the inset are the total number of Gaussians that provide the maximum classification accuracy across shuffles for this exemplar subject. Color assignment is only for display purposes, it does not carry any information used by the classifier. The total number of maps for each of the two conditions, *Q*1 and *Q*2, are shown in the inset. The thick yellow line on the *x*-axis indicates the temporal periods of difference between conditions. (b) Template maps (back dots localize the maximum and minimum voltage values), previously normalized by Global Field Power. Frame colors correspond to the posterior probabilities shown in panel a. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

posterior probabilities across trials revealed a structured modulation in the post-stimulus period for all subjects (see Fig. 7 for the results of an exemplar subject and Supplemental Fig. 2 for results from all other subjects).

The Bayes factors revealed differences between initial and repeated stimuli starting at 270 ms post-stimulus onset and persisting up to 350 ms for the exemplar subject shown in Fig. 7. Later differences were also observed starting around 455 ms. The corresponding periods of difference for the rest of the subjects are shown in Supplemental Fig. 2. These periods of difference overlap with earlier VEP findings not only of the same dataset [28], but also with many prior studies [49]. The posterior probabilities shown in Fig. 7 are the ones that yielded this difference and were then used for classification on the test dataset. The thick yellow line on the *x*-axis in Fig. 7 shows the time intervals for which the difference was found in at least 8 of the 10 shuffles.

4.2.3. Classification results

The classification based on AUC with *single-trial* test data was not significantly different from chance. However, it should be noted that in contrast to the above Checkerboard Experiment, in which differences were predicted from known retinotopic representations, topographic differences between task conditions are likely to be more subtle. An alternative approach, often applied in BCI, is to improve the signal-to-noise ratio by averaging a small number of trials and then performing the classification [50–52]. When we applied this strategy here, using averages of five trials instead of single responses, we found an average across subjects AUC of 0.74 and for three out of four subjects classification rates were significantly above chance level (unpaired *t*-test, $p < 0.05$). The averaged ROC curves are shown in the inset of Fig. 8 (mean $\pm$ s.e.m.). For display purposes, the ROC curves are shown after collapsing the results across shuffles. The absolute value of classification accuracy was 0.63 on average across subjects and was found above chance levels for all subjects (unpaired *t*-test, $p < 0.05$) but the second (S2 in Supplemental Material 2).

The average AUC in the validation dataset (including 20 trials per condition and per subject) was on average 0.61. This average is computed across three subjects only, because for one of the subjects (S2 in Supplemental Material 2) we did not have enough artifact-free trials to be considered as a separate dataset for validation.

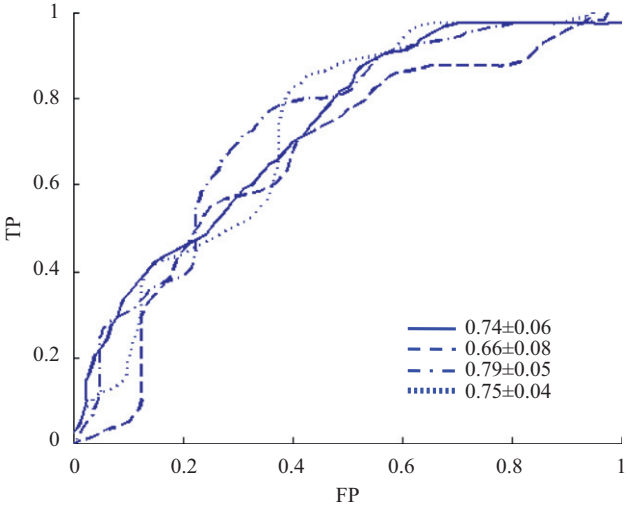

Fig. 8. ROC curve area for each of the subjects collapsed across shuffles (FP and TP stand for rate of false positive and rate of true positive respectively) for the Priming Experiment. Each of these ROC curves result from classifying sub-averages of 5 trials corresponding to initial versus repeated presentation of identical visual objects. Model selection (number of Gaussians in the two GMM models) was optimized for each of the subjects. Mean values across shuffles of the average AUC are shown (mean $\pm$ s.e.m.) from subject S1 to S4. All these values are significantly above chance but that of subject S2 (*t*-test; $p > 0.06$). The exemplar subject shown in Fig. 7 corresponds to S4 for both upper and lower visual field (the others are shown in the Supplemental Material 2).

The robustness of these results and the effect of the expectation-maximization initialization on the GMM models were also tested, as in the Checkerboard Experiment. We recomputed the GMM models for the exemplar subject and then selected the total number of Gaussians using ten different initializations on the *k*-means [48]. Across the initializations and shuffles the AUC was on average 0.71 ± 0.02 ranging from 0.63 up to 0.78 for different initializations. The previously reported AUC value for this subject was 0.75.

4.2.4. Classification based on average ERP model

In the case where we re-trained the models using the average ERP data and classified the averages of five trials of the test

Table 2
ROC curve areas for all the Priming Experiment based on the single-trial models and the average ERP models. The first column displays the ROC curve areas ($\pm$ standard error) in the case where the classification is based on models computed using all the single-trial data, while the second shows the results using the average ERPs. The relative change column refers to the relative difference in the ROC curve area, when computed using the average ERP model in comparison with the case where we use the single-trial models (asterisks indicates those differences that were significant; paired t-test, $p<0.05$).

	Single-trial	Average ERP	Relative change (%)
S1	0.74 ± 0.06	0.65 ± 0.06	-13
S2	0.66 ± 0.08	0.73 ± 0.05	$+10$
S3	0.79 ± 0.05	0.72 ± 0.04	$-9*$
S4	0.75 ± 0.04	0.67 ± 0.06	$-11*$

datasets, there was a significant drop in AUC for two out of four subjects (S3 and S4 in Table 2; paired t-test, $p<0.002$). There was no significant difference for the other two subjects ($p \geq 0.35$). When using the models trained on single-trial data, the average AUC was 0.73 for these two subjects, while in the case of using the models trained on average ERPs the average AUC was 0.70, or had a decrease of 6%.

4.2.5. Comparison between template maps and average ERP maps

As explained for the Checkerboard Experiment, 'reconstructed ERP' maps can be generated from a weighted average of the template maps according to their posterior probabilities at every time-point. These maps, compared to the average ERP topographies, yielded a high spatial correlation in the entire post-stimulus period (ranging from 0.73 to 0.86 for three out of four subjects for both conditions and 0.58 to 0.65 for the fourth), which resulted in Global Explained Variance values ranging from 0.66 to 0.85 for three out of the four subjects and from 0.61 to 0.71 for the fourth subject (see Supplemental Material 4 for an overview of these maps).

5. Discussion

We presented a novel approach for classifying single-trials (or sets of a few single trials) of ERPs based on voltage topographies with minimal *a priori* constraints. In two independent VEP datasets we provided evidence that voltage topographies can be used to accurately classify differences between experimental conditions, both when topographic differences are expected based on retinotopic functional anatomy (Checkerboard Experiment), and when topographic differences are evaluated in a more exploratory manner (Priming Experiment).

The main advantage of this approach is that topographic features are readily interpretable in terms of their neurophysiologic basis; a change in their presence across trials and between conditions forcibly reflects a change in the underlying generator configurations. Moreover, as these maps were previously normalized by the overall strength of the field potential on the scalp (GFP), this difference cannot be due to active generators with the same spatial configuration that modulate their activity strength. However, it is noteworthy that taking into account GFP modulation in addition to topographic information would potentially boost our classification accuracy, because we would take advantage of some possible strength modulation effects. On the other hand, this would limit our results' interpretability as we would not know whether the difference between the experimental conditions is due to a different source configuration and/or to a modulation in the strength of the same sources. In particular in the Priming Experiment, even if we did not exploit any repetition-related reductions in neural activity [49] – i.e. modulations in response strength – we nonetheless obtained above-chance classification performance in three of the four subjects.

By using a cross-validation, our procedure assures that above-chance classification performance is not the result of over-fitting. Training and testing were performed on separate splits of the data, and the ten test datasets were obtained by non-overlapping (independent) dataset shuffles. However a final conclusion on the algorithm's performance could be obtained by applying the classification on a validation dataset, independent from the one used for model selection. As expected, in both experiments we obtained a lower performance in the validation dataset with respect to what was obtained in the test datasets. Nevertheless, we obtained on average above-chance classification accuracy in both experiments.

The cross-validation offers a general solution to optimize the number of template maps in the GMM models of the two conditions for each subject. This optimization was reached so as to maximize the discrimination power of the classifier and not to best fit each of the conditions separately. Nevertheless, even if we did not take into account the GEV in our selection criteria, the GEV of the two datasets was on average 64% for the Checkerboard Experiment when contrasting the upper quadrants and the same value on average across subjects when contrasting the lower quadrants. The GEV was also high in the Priming Experiment yielding an average of 76% across subjects and conditions.

The latencies at which we obtained a reliable difference between conditions were estimated for each training dataset for ten splits of the data. By this procedure it was possible to determine the robustness of these differences. Indeed, across the ten shuffles and in both experiments, we observed a high level of consistency in the temporal periods over which the two conditions differed. Given this consistency, we could retain as reliable those data-points over which a difference between conditions was estimated in at least eight of the ten shuffles.

These periods of temporal differences were based on computing the Bayes factor at the same latency for the two conditions (see Sections 2.1 and 2.1.3). However, the single-trial classification exploits temporally unlocked contributions in two steps of the implementation. First, the models are computed using the ensemble of trials without taking into account their specific latency. Second, in the test phase the posterior probabilities are computed allowing for a temporal jitter ($u \ll H$) in time points (footnote 2); that is to say we consider the possibility that the difference found in the training dataset between the two conditions could appear shifted in time. Indeed, introducing this jitter increases the classification accuracy.

5.1. Comparison with classification based on average ERP model

We evaluated the benefit of using single-trial topographic classifiers by comparing classification results based on single-trial models with those based on the average ERP response. We expected higher classification performance for models based on the single trial ERP responses because those exploit both signals that are time-locked to stimulus onset, and those that are not. This was indeed our empirical observation. We obtained a significantly higher AUC for three out of four subjects in the Checkerboard Experiment and for two out of four subjects in the Priming Experiment (Tables 1 and 2). Importantly, we never obtained a significantly lower AUC when training the model on the single-trial ERP. This confirms that single-trial data contains useful information for stimulus classification despite the fact that average ERP has obviously higher SNR. This shows that non stimulus-locked activity plays a significant role in elementary sensory processes even in simple paradigms where minimal cognitive factors are involved.

5.2. Comparison with average ERPs

The classifier extracted features that overlapped with the averaged ERP topographies in both experiments. In the Checkerboard Experiment, the features resembled scalp topographies that may be expected based on retinotopic functional anatomy [45,46]. In particular, we found for all the subjects early differences between conditions starting at around the latency of the C1 component extending up to the N1 latency (depending on the subject). At these latencies, the template maps exhibited the correct polarity and the appropriate contra-laterality with respect to those found in average data, as can be seen by comparing the template maps in Fig. 4 to the average ERP maps in Fig. 3. The average spatial correlation between the template and the average maps was 0.74 during the post-stimulus period, across subjects and conditions.

For the Priming Experiment, we compared our results with the averaged ERP results [28]. Although the latencies of our effects fall within those observed at the group-level, this previous study also revealed an early topographic effect starting at 36 ms post-stimulus onset. Possible explanations for the discrepancy between the two analyses are that the current one used only a subset of the subjects, and that the original dataset was down-sampled to fewer electrodes (i.e. nearly half). The consequent lower spatial resolution is possibly hiding some subtle topographic differences between conditions. An important direction for future research, particularly for clinical applications, will be to determine the sensitivity of the methods to the spatial sampling of the electric field at the scalp, as this will impact the suggested number and distribution of scalp electrodes.

5.3. Relation to other classification methods

EEG single-trial classification is the focus of a continuous effort for developing BCI machines [53,54]. While we can definitely benefit from a longstanding tradition in this domain, it is worth noting that BCI puts its emphasis on online classification accuracy, whereas the application to ERP studies and more generally to neuroscientific research questions must also emphasize neurophysiologic interpretability. Whereas high classification accuracy is required for BCI applications, smaller accuracy is acceptable for neuroscientific aims. Future application of the present approach for BCI purposes remains to be explored especially in relation to experimental designs that can be used to control external devices for online applications.

Our classification algorithm relates to existing data reduction approaches, for example PCA and ICA (see also [55]) in that it extracts a limited number of prototypical voltage topographies. However the purpose of PCA and ICA is to reduce redundant information by finding an optimal set of basis functions each accounting for an orthogonal/independent component, respectively. To this aim PCA imposes orthogonality and ICA imposes independency of the EEG components, in the temporal or in the spatial domain. As a result of this decomposition it is often possible to explain a large amount of variance while retaining only few template maps [56,57]. In parallel, the mathematical constraints that are imposed to estimate these template maps do not assure an optimal separation of the topographies reflecting different underlying sources. One typical example is given by sources whose activity is reflected on the scalp as Gaussians with overlapping distribution (Supplemental Material 3). In this example, PCA finds the directions of highest variance of only one Gaussian distribution, therefore failing to separate the two Gaussians. In contrast, ICA cannot separate Gaussian distributions because by definition it extracts directions with the least Gaussian distribution. In this case, modeling the dataset with a GMM with two Gaussians, we can easily disentangle the two template topographies and their periods of activations.

5.4. Future directions

Presently, we have shown how to optimize the classification at single subject level. It is of course interesting being able to quantify the degree of similarity between subjects, particularly when one wants to carry out the analysis at group-level. Therefore, we plan to generalize our classification method based on topographies extracted at single-trial level from several subjects [25]. Being able to analyze ERPs at the single subject as well as group level provides a novel tool to develop ERP classification in normal electrophysiological responses versus single cases that cannot be considered part of any cohort of subjects [58,59]. This aspect promises to have a strong impact on clinical studies.

From a more methodological standpoint, two main directions can be further investigated: alternative approaches for model estimation and different criteria for model selection. For model estimation in this study, we proposed a classical expectation-maximization algorithm. This approach has two main limitations. It can get caught in local maxima and it is also strongly dependent upon its initialization. Here, in order to overcome these limitations we used a k-means algorithm to get an initial estimation of the model's means, but the k-means also needs an initialization which might affect its final output. We quantified the degree to which different initializations of the k-means affect our final results and we demonstrated that classification accuracy remains relatively stable (accuracy range for different initializations was well within the standard error across test datasets). We also plan to further explore different techniques for selecting the best initialization, such as multiple initialization techniques or tracking the possible expectation-maximization trajectories [48,60].

In alternative to expectation-maximization, Bayesian techniques (e.g. Variational Bayes) [33] and a more advanced version of the expectation-maximization, (e.g. 'splitting and merging' Gaussians) [61] can also be explored. These approaches have been demonstrated to be less prone to over-fitting, so we expect to further improve the GMM estimation and classification accuracy.

As for model selection, GMM modeling requires *a priori* information about the total number of classes. In order to perform model selection, we cross-validated across the ten shuffles and chose the model that best discriminated between experimental conditions. Alternatively, parameters' selection can be based on complexity criteria like Bayesian information criterion [62], Akaike information criterion [63] or Minimum Descriptor Length [64] and Minimum Message Length [65]. Such criteria essentially quantify the prediction error, which depends on the actual training error and the complexity of the model, thus preventing over-fitting. Although a comparison between these different approaches is foreseen, so far we have preferred to adopt an empirical approach for model selection. These other strategies can become crucial when a limited amount of trials will prevent training and testing the model on separate splits of the data.

So far, at every time-latency we have only used a single Gaussian for classifying, but this can be easily extended to more. Using more classes or even the whole mixture might improve results, but this needs to be further investigated. Finally, in the current study we only considered contrasts between pairs of experimental conditions, e.g. the upper and lower visual fields in the Checkerboard Experiment. However, this approach can readily be extended to deal with more than two conditions in a multiclass approach.

Acknowledgments

The Swiss National Science Foundation provided financial support (Grants #K-33K1_122518/1 to MDL and 310030B_133136 to MMM).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.patcog.2011.04.007.

References

[1] S. Makeig, M. Westerfield, T.P. Jung, S. Enghoff, J. Townsend, E. Courchesne, T.J. Sejnowski, Dynamic brain sources of visual evoked responses, Science 295 (2002) 690–694.

[2] R.Q. Quiroga, H. Garcia, Single trial event-related potentials with wavelet denoising, Clin. Neurophysiol. 114 (2003) 376–390.

[3] S.L. Gonzalez Andino, M.M. Murray, J.J. Foxe, R.G. de Peralta Menendez, How single-trial electrical neuroimaging contributes to multisensory research, Exp. Brain Res. 166 (2005) 298–304.

[4] K.H. Knuth, A.S. Shah, W.A. Truccolo, M. Ding, S.L. Bressler, C.E. Schroeder, Differentially variable component analysis: identifying multiple evoked components using trial-to-trial variability, J. Neurophysiol. 95 (2006) 3257–3276.

[5] R. Ratcliff, M.G. Philiastides, P. Sajda, Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG, Proc. Natl. Acad. Sci. USA 106 (16) (2009) 6539–6544.

[6] R.B. Mars, S. Debener, T.E. Gladwin, L.M. Harrison, P. Haggard, J.C. Rothwell, S. Bestmann, Trial-by-trial fluctuations in the event-related electroencephalogam reflect dynamic changes in the degree of surprise, J. Neurosci. 28 (47) (2009) 12539–12545.

[7] M.L.A. Jongsma, R.Q. Quiroga, C.M. van Rijn, Rhytmic training decreases latency-jitter of omission evoked potentials (OEPs) in humans, Neurosci. Lett. 355 (2004) 189–192.

[8] M.L.A. Jongsma, T. Eichele, C.M.V. Van Rijn, A.M.L. Coenen, K. Hugdahl, H. Nordby, R.Q. Quiroga, Tracking pattern learning with single-trial event-related potentials, Clin. Neurophysiol. 117 (2006) 1957–1973.

[9] J.D. Haynes, G. Rees, Decoding mental states from brain activity in humans, Nat. Rev. Neurosci. 7 (2006) 523–534.

[10] M.G. Philiastides, P. Sajda, Temporal characterization of the neural correlates of perceptual decision making in the human brain, Cereb. Cortex 16 (4) (2006) 509–518.

[11] M.G. Philiastides, G. Biele, N. Vavatzanidis, P. Kazzer, H.R. Heekeren, Temporal dynamics of prediction error processing during reward-based decision making, Neuroimage 53 (1) (2010) 221–232.

[12] A. Tusche, S. Bode, J.D. Haynes, Neural responses to unattended products predict later consumer choices, J. Neurosci. 30 (23) (2010) 8024–8031.

[13] M. Bles, J.D. Haynes, Detecting concealed information using brain-imaging technology, Neurocase 14 (1) (2008) 82–92.

[14] E. Formisano, F. De Martino, M. Bonte, R. Goebel, "Who" is saying "what"? Brain-based decoding of human voice and speech, Science 322 (5903) (2008) 970–973.

[15] T. Ethofer, D. Van De Ville, K. Scherer, P. Vuilleumier, Decoding of emotional information in voice-sensitive cortices, Curr. Biol. 19 (12) (2009) 1028–1033.

[16] K. Meyer, J.T. Kaplan, R. Essex, C. Webber, H. Damasio, A. Damasio, Predicting visual stimuli on te basis of activity in auditory cortices, Nat. Neurosci. 13 (6) (2010) 667–668.

[17] M. Misaki, Y. Kim, P.A. Bandettini, N. Kriegeskorte, Comparison of multivariate classifiers and response normalizations for pattern-information fMRI, Neuromage 53 (1) (2010) 103–118.

[18] D. Lehmann, Principles of spatial analysis, in: A.S. Gevins, A. Remond (Eds.), Handbook of Electroencephalography and Clinical Neurophysiology, Vol. 1: Methods of Analysis of Brain Electrical and Magnetic Signals, Elsevier, Amsterdam, 1987, pp. 309–354.

[19] C.M. Michel, G. Thut, S. Morand, A. Khateb, A.J. Pegna, R. Grave de Peralta, S. Gonzales, M. Seeck, T. Landis, Electric source imaging of human brain functions, Brain Res. Rev. 36 (2001) 108–118.

[20] C.M. Michel, M.M. Murray, G. Lantz, S. Gonzalez, R. Grave de Peralta, EEG source imaging, Clin.Neurophysiol. 115 (2004) 2195–2222.

[21] M.M. Murray, D. Brunet, C.M. Michel, Topographic ERP analyses: a step-by-step tutorial review, Brain Topogr. 20 (2008) 249–264.

[22] J. Lefèvre, S. Baillet, Optical flow approaches to the identification of brain dynamics, Hum. Brain Mapp. 30 (6) (2009) 1887–1897.

[23] D. Van de Ville, J. Britz, C.M. Michel, EEG microstate sequences in healthy humans at rest reveal scale-free dynamics, Proc. Natl. Acad. Sci. USA 107 (42) (2010) 18179–18184.

[24] D. Brunet, M.M. Murray, C.M. Michel, Spatiotemporal analysis of multichannel EEG: CARTOOL, Comput. Intell Neurosci. (2011) 813870.

[25] M. De Lucia, C.M. Michel, S. Clarke, M.M. Murray, 2007a. Single-trial topographic analysis of human EEG: a new 'image' of event-related potentials, in: Proceedings Information Technology Applications in Biomedicine 2007.

[26] M. De Lucia, C.M. Michel, S. Clarke, M.M. Murray, Single subject EEG analysis based on topographic information, Int. J. Bioelectromagn. 9 (2007) 168–171.

[27] M.M. Murray, C.M. Michel, R. Grave de Peralta, S. Ortigue, D. Brunet, S. Gonzalez Andino, A. Schnider, Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging, Neuroimage 21 (1) (2004) 125–135.

[28] M.M. Murray, M. De Lucia, D. Brunet, C.M. Michel, Principles of topographic analyses of electrical neuroimaging, in: T.C. Handy (Ed.), Event-Related Potentials II: Advances in ERP, EEG, & MEG Analysis, MIT Press, 2009.

[29] G. Plomp, C.M. Michel, M.H. Herzog, Electrical source dynamics in three functional localizer paradigms, Neuroimage 53 (1) (2010) 257–267.

[30] D. Lehmann, W. Skrandies, Reference-free identification of components of checkerboard-evoked multichannel potential fields, Electroencephalogr. Clin. Neurophysiol. 48 (1980) 609–621.

[31] T. Koenig, L.A. Melie-García, Method to determine the presence of averaged event-related fields using randomization tests, Brain Topogr. 23 (3) (2010) 233–242.

[32] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. R. Statist. Soc. B 39 (1977) 1–38.

[33] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[34] J. Millán, R. del, J. Mouriño, M. Franzé, F. Cincotti, M. Varsta, J. Heikkonen, F.A. Babiloni, Local neural classifier for the recognition of EEG patterns associated to mental tasks, IEEE Trans. Neural Networks 13 (2002) 678–686.

[35] R. Chavarriaga, J. Millán, R. del, Learning from EEG error-related potentials in noninvasive brain–computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 18 (2010) 381–388.

[36] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. R. Statist. Soc. B 58 (1996) 155–176.

[37] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Data Mining, Elements and Prediction, Springer, 2001, pp. 399–405.

[38] W.D. Penny, K.E. Stephan, A. Mechelli, K.J. Friston, Comparing dynamic causal models, Neuroimage 22 (3) (2004) 1157–1172.

[39] A.E. Raftery, Bayesian model selection in social research, Sociol. Methodol. 25 (1995) 111–163.

[40] N.A. Macmillan, D.C. Creelman, Detection Theory: A User's Guide, second edition, Lawrence Erlbaum Associates, New Jersey, 2005.

[41] R.C. Oldfield, The assessment and analysis of handedness: the Edinburgh Inventory, Neuropsychologia 9 (1971) 97–113.

[42] J.G. Snodgrass, M. Vanderwart, A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity, J. Exp. Psychol.: Hum. Learning Mem. 6 (1980) 174–215.

[43] F. Perrin, J. Pernier, O. Bertrand, M.H. Giard, J.F. Echallier, Mapping of scalp potentials by surface spline interpolation, Electroencephalogr. Clin Neurophysiol. 66 (1987) 75–81.

[44] N. Kriegeskorte, W.K. Simmons, P.S.F. Bellgowan, C.I. Baker, Circular analysis in systems neuroscience: the dangers of double dipping, Nat. Neurosci. 12 (2009) 535–540.

[45] V.P. Clark, S. Fan, S.A. Hillyard, Identification of early visual evoked potential generators by retinotopic and topographic analyses, Hum. Brain Mapp. 2 (1995) 170–187.

[46] F. Di Russo, S. Pitzalis, T. Aprile, G. Spitoni, F. Patria, A. Stella, D. Spinelli, S.A. Hillyard, Spatiotemporal analysis of the cortical sources of the steady-state visual evoked potential, Hum. Brain Mapp. 28 (4) (2007) 323–334.

[47] J.J. Foxe, E.C. Strugstad, P. Sehatpour, S. Molholm, W. Pasieka, C.E. Schroeder, M.E. McCourt, Parvocellular and magnocellular contributions to the initial generators of the visual evoked potential: high-density electrical mapping of the "C1" component, Brain Topogr. 21 (1) (2008) 11–21.

[48] M.A.T. Figueiredo, Unsupervised Learning of finite mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 381–396.

[49] K. Grill-Spector, R. Henson, A. Martin, Repetition and the brain: neural models of stimulus-specific effects, Trends Cogn. Sci. 10 (1) (2006) 14–23.

[50] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain-computer interfaces for communication and control, Clin. Neurophysiol. 113 (6) (2002) 767–791.

[51] F. Guo, B. Hong, X. Gao, S. Gao, A brain–computer interface using motion-onset visual evoked potential, J. Neural Eng. 5 (2008) (2008) 477–485.

[52] A. Lenhardt, M. Kaper, H.J. Ritter, An adaptive P300-based online brain–computer interface, IEEE Trans. Neural Syst. Rehabil. Eng. 16 (2) (2008) 121–130.

[53] F. Galán, M. Nuttin, E. Lew, P.W. Ferrez, G. Vanacker, J. Philips, R. Millán Jdel, A brain-actuated wheelchair: asynchronous and non-invasive Brain-computer interfaces for continuous control of robots, Clin. Neurophysiol. 119 (9) (2008) 2159–2169.

[54] C.S. Da Salla, H. Kambara, M. Sato, Y. Koike, Single-trial classification of vowel speech imagery using common spatial patterns, Neural Netw. 22 (9) (2009) 1334–1339.

[55] M. De Lucia, C.M. Michel, M.M. Murray, Comparing ICA-based and single-trial topographic ERP analyses, Brain Topogr. 23 (2) (2010) 119–127.

[56] K. Lugger, D. Flotzinger, A. Schlögl, M. Pregenzer, G. Pfurtscheller, Feature extraction for on-line EEG classification using principal components and linear discriminants, Med. Biol. Eng. Comput. 36 (3) (1998) 309–314.

[57] S. Makeig, M. Westerfield, T.P. Jung, J. Covington, J. Townsend, T.J. Sejnowski, E. Courchesne, Functionally independent components of the late positive event-related potential during visual spatial attention, J. Neurosci. 19 (7) (1999) 2665–2680.

[58] S. Barcellona-Lehmann, S. Morand, C. Bindschaedler, L. Nahum, D. Gabriel, A. Schnider, Abnormal cortical network activation in human amnesia: a high-resolution evoked potential study, Brain Topogr. 23 (1) (2010) 72–81.

[59] M. Laganaro, C. Perret, Comparing electrophysiological correlates of word production in immediate and delayed naming through the analysis of word age of acquisition effects, Brain Topogr. 24 (1) (2011) 19–29.

[60] C. Biernacki, Initializing EM using the properties of its trajectories in Gaussian mixtures, Stat. Comput. 14 (3) (2004) 267–279.
[61] N. Ueda, R. Nakano, Z. Gharamani, G.E. Hinton, Split and merge EM algorithm for improving gaussian mixture density estimates, J. VLSI Signal Process. Syst. 26 (2000) 133–140.
[62] G.E. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[63] H. Akaike, 1973. Information theory and an extension of the maximum likelihood principle, in: Proceedings of the Second International Symposium on Infromation Theory, pp. 267–281.
[64] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.
[65] C. Wallace, O. Freeman, Estimation and inference via compact coding, J. R. Stat. Soc. 49 (1987) 241–252.

Athina Tzovara received her Electrical and Computer Engineering degree in 2009 from National Technical University of Athens, Greece. She is currently working towards the Ph.D. degree in the Functional Electrical Neuroimaging Lab of the University of Lausanne, Switzerland. Her interests are in machine learning applications for brain decoding based on electroencephalography.

Micah M. Murray received his B.A. from The Johns Hopkins University and his Ph.D. in neuroscience from Albert Einstein College of Medicine of Yeshiva University. Since 2003 he has held a position within the Department of Clinical Neurosciences and Department of Radiology at the University Hospital of Lausanne, Switzerland. Currently, he is an associate professor within these departments, adjunct associate professor at Vanderbilt University, as well as associate director of the EEG Brain Mapping Core of the Center for Biomedical Imaging in Lausanne, Switzerland. He has authored more than 90 articles and book chapters. His group's research primarily focuses on multisensory interactions, object recognition, learning and plasticity, EEG/fMRI methodological developments, and systems/cognitive neuroscience in general. Research in his group combines psychophysics, EEG, fMRI, and TMS in healthy and clinical populations.

Gijs Plomp received a Master's Degree in cognitive and biological psychology from the university of Amsterdam and a Ph.D. from the University of Sunderland in 2005 for his work on visual completion phenomena, done at the Riken Brain Science Institute in Tokyo. He was a post-doc at the Ecole Polytechnique Fédérale de Lausanne (EPFL, Switzerland), and then received an Ambizione grant from the Swiss National Fund to work at the University of Geneva. His main research interest is in the dynamics of visual processes in the brain.

Michael H. Herzog was born in Erlangen, Germany, on October 11, 1964. He received the B.A. degree in Mathematics from the University of Erlangen, Germany in 1988 and the M.A. (Diplom) degree in mathematics, the M.A. degree in philosophy, and the Ph.D. degree in biology, all from the University of Tubingen, Germany, in 1992, 1993 and 1996, respectively. From 1998 to 1999, he was a Postdoctoral Researcher at the California Institute of Technology, Pasadena, investigating the characteristics of temporal processing and feature integration. From 1999 to 2004, he was a Senior Researcher at the Section of Human Neurobiology at the University of Bremen, Germany, and was a Leader of a research project at the Center of Excellence 517 BNeurocognition of the DFG (German Research Council). He also held a temporary Professorship for Neurobiopsychology at the University of Osnabrück, Germany for one year. Currently, he is a Professor of Neuroscience and Psychophysics at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. He is currently the Head of the Laboratory of Psychophysics at the Brain Mind Institute of EPFL. His research interests are in brain mechanisms of visual processing in humans. Prof. Herzog is a member of the Vision Sciences Society (VSS).

Christoph M. Michel is Professor for Clinical Neuroscience at the Medical Faculty of the University of Geneva. He is head of the Functional Brain Mapping Laboratory at the Neurology Clinic and the Fundamental Neurosience Department and Director of the EEG core of the Lemanic Biomedical Imaging Center. He has authored over 160 articles and book chapters and is co-author of the book "Electrical Neuroimaging".

The principal research interest of the lab is the organization and the dynamics of the large-scale neuronal networks of the human brain that characterize mental functions, and the understanding of disturbances of these networks in patients with brain dysfunctions. Electromagnetic imaging based on high-resolution EEG is the main instrument to study these questions and is combined with fMRI, diffusion imaging and TMS. The clinical research focuses on non-invasive localization of epileptic networks and of eloquent cortex in the context of presurgical epilepsy evaluation.

Marzia De Lucia received her Ph.D. in Physics in 2004 from the University La Sapienza, Rome, Italy. She then joined as post-doc the Bioengineering and Medical Physics Department, University College London, United Kingdom. Currently she is lecturer within the faculty of Biology and Medicine, University of Lausanne, Switzerland. Her main research focuses are categorical discrimination and plasticity in human audition using electrical neuroimaging and decoding algorithms applied to electroencephalography and intracranial recordings.