Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals

Jukka-Pekka Kauppi[†a,b,*], Melih Kandemir[†f,d], Veli-Matti Saarinen[b], Lotta Hirvenkari[b], Lauri Parkkonen[b,f], Arto Klami[a], Riitta Hari[b,e], Samuel Kaski[a,f,*]

[a] *Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland*
[b] *Brain Research Unit, O.V. Lounasmaa Laboratory, School of Science, Aalto University, Espoo, Finland*
[c] *Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, Espoo, Finland*
[d] *Heidelberg University HCI/IWR, Heidelberg, Germany*
[e] *MEG Core, Aalto NeuroImaging, Aalto University, Espoo, Finland*
[f] *Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland*

Abstract

We hypothesize that brain activity can be used to control future information retrieval systems. To this end, we conducted a feasibility study on predicting the relevance of visual objects from brain activity. We analyze both magnetoencephalographic (MEG) and gaze signals from nine subjects who were viewing image collages, a subset of which was relevant to a predetermined task. We report three findings: i) The relevance of an image a subject looks at can be decoded from MEG signals with performance significantly better than chance, ii) fusion of gaze-based and MEG-based classifiers significantly improves the prediction performance compared to using either signal alone, iii) non-linear classification of the MEG signals using Gaussian process classifiers outperforms linear classification. These findings break new ground for building brain activity based interactive image retrieval systems, as well as for systems utilizing

*Corresponding author. †Authors contributed equally to this work.
Email addresses: jukka-pekka.kauppi@helsinki.fi (Jukka-Pekka Kauppi†), melih.kandemir@iwr.uni-heidelberg.de (Melih Kandemir†), veli-m@neuro.hut.fi (Veli-Matti Saarinen), lotta.hirvenkari@aalto.fi (Lotta Hirvenkari), lauri.parkkonen@aalto.fi (Lauri Parkkonen), aklami@cs.helsinki.fi (Arto Klami), riitta.hari@aalto.fi (Riitta Hari), samuel.kaski@hiit.fi (Samuel Kaski)

feedback both from brain activity and eye movements.

Keywords: Bayesian classification, image relevance, implicit relevance feedback, information retrieval, magnetoencephalography, gaze signal

1. Introduction

Interactive information retrieval (IIR) refers to a process where a user searches for information in an iterative manner. A typical interactive search interface provides a set of search results at a time, collects *relevance feedback* for the provided items, and then proceeds to provide a new result set that hopefully is more relevant[1] for the search needs of the user (Ruthven, 2008).

Most IIR systems collect the relevance feedback by explicitly asking for it from the user. To relieve the user from this burden, implicit relevance feedback can be inferred from various signals from the user, including behavioral features such as mouse movements (Claypool et al., 2001b), gaze patterns (Klami et al., 2008; Puolamäki et al., 2005; Salojärvi et al., 2005; Hardoon and Pasupa, 2010; Ajanki et al., 2009), and physiological measurements such as skin conductance (Soleymani et al., 2008). It is to be expected that future IIR systems will routinely use these kinds of additional information sources as soon as the necessary measurement techniques are widely available; they have been demonstrated to provide useful information that should not be ignored. However, the information available in the physiological signals measured outside the brain is limited. For example, gaze trackers are effective in revealing where the user looks at, and facial expressions provide clues on the users' emotional responses (Arapakis et al., 2009), but it is unlikely that more complex behavior could be accurately decoded without direct measurements of brain activity.

In this work, we provide one of the first steps in exploring whether rhythmic

[1] *Relevance* is a highly important concept in information science and has been utilized in practical information retrieval systems as well (Saracevic, 2007). The concept of relevance is abstract and has multiple interpretations. We define relevance as the *importance of a stimulus in fulfilling the goal of a given cognitive task.*

brain activity could provide richer feedback information not available in other physiological signals besides brain signals. More specifically, we asked subjects to find, by free visual exploration, the most relevant images in collages of natural scenes, and hypothesized that relevance of an image can be decoded from magnetoencephalographic (MEG) (Hari and Salmelin, 2012) signals of the subjects. We also hypothesized that the prediction performance based on the combination of MEG and gaze signals improves decoding performance compared to just using gaze signals. MEG provides dynamic information of brain function with a millisecond-scale temporal resolution and is therefore well-suited to our study.

Multivariate classification of functional neuroimaging signals has been extensively studied in the context of brain–computer interfaces (BCIs), which allow direct communication between the user's brain and an external device, by distinguishing between brain signatures of the users' intentions (Wolpaw et al., 2002). For instance, Pohlmeyer et al. (2011) presented a closed-loop BCI system, based on electroencephalography (EEG), for image search from large databases. In contrast to BCI studies, we do not provide the user direct means of controlling the search, but instead only infer the relevance of the images passively. Thus the user can focus on the search task itself in a natural way, instead of needing to perform additional tasks, such as motor imagery, required by many BCI techniques. In general, our free-viewing setup makes our study quite different from BCI and brain-function decoding studies (Haynes and Rees, 2006) that typically use highly controlled stimuli and viewing conditions. Recently, Eugster et al. (2014) showed that the relevance of words can be predicted from EEG signals with written-text stimuli, and Moshfeghi et al. (2013) decoded binary relevance information of an image from fMRI. Both these studies used block design to control stimulus presentation. In contrast to these studies, we used a free-viewing setup and thus identified the relevance in conditions much closer to real usage. Free-viewing setups have been previously used for EEG-based visual target detection, but the objectives and experimental designs of these studies differ considerably from our study. For instance, Jangraw et al. (2014) built a hybrid BCI system for navigation in 3D environment, and Dias et al.

(2013) predicted, based on EEG signals acquired prior to possible fixation to the target, whether the subject would *detect* or *miss* the small target object (would or would not fixate to it).

1.1. Relevance feedback

In IIR systems, the actual retrieval of the new results that takes into account the relevance feedback can be performed in a multitude of ways. In this work, we exclusively focus on the process of *obtaining* the feedback. Relevance feedback can be given in several ways, often dichotomized to *explicit* and *implicit* feedback. Explicit feedback refers to systems that explicitly ask the user to label the images as relevant or non-relevant (Claypool et al., 2001a; Fox et al., 2005; Joachims et al., 2005) or which assume that the choices made by the user can be directly interpreted as positive feedback (Datta et al., 2008; Lew et al., 2006). A user selecting one image out of several can be seen as giving partial feedback, indicating that the particular image is more relevant than the others. Implicit feedback, in turn, refers to techniques that attempt to automatically infer the relevance by monitoring the user. The motivation for using implicit feedback is in relieving the user from the laborious task of providing the feedback. Even though the estimated relevance scores are typically not perfect characterizations of the user's needs, they can still be used for guiding the search (Buscher et al., 2012) and the possibility of obtaining perfect coverage (relevance feedback for all possible items) without any explicit requirements for the users is intriguing.

Although implicit feedback is typically obtained by analyzing explicit behavioral actions (tracking mouse movements, scrolling, link clicks, etc), the more relevant perspective for this work is provided by studies inferring the feedback from physiological signals recorded from users. The most common information source for this kind of monitoring has been gaze direction, which is easy to record without bothering the user. For instance, gaze-related features have been used for inferring the relevance of items in either textual (Puolamäki et al., 2005) or visual (Klami et al., 2008) retrieval tasks. Also other forms of physiological signals have been used; Soleymani et al. (2008) used galvanic skin responses,

heart-rate and skin temperature for estimating users' interest in movie scenes, and Moshfeghi and Jose (2013) combined behavioral and physiological signals, such as skin temperature, facial expressions, EEG, heart rate, motion data and dwell-time, for obtaining feedback for video retrieval.

In practical terms, the task of inferring implicit relevance feedback is often formulated as binary classification ("relevant" vs "non-relevant") for each item in a result set, either producing as output the estimated classes themselves or the probability of relevance. Hence, the computational tools for implicit relevance prediction are classifiers that take as inputs a feature representation for each item and then produce as the output the probabilities of the two classes. Alternatively, relevance feedback could also be provided by ranking the items according to the estimated relevance, for example by simply sorting them according to the probabilities of relevance, or by directly estimating the ranks (Huang et al., 2010; Lan et al., 2012).

2. Materials

2.1. Subjects

Nine volunteers (5 male, 4 female; 21—30 years, mean 25.8 years) with normal vision participated in the experiment. Before starting the experiment, the course of the study was explained, and the participants gave their written informed consent. The recordings had a prior approval by the Ethics Committee of the Hospital District of Helsinki and Uusimaa.

2.2. Eye Tracking

Eye movements were measured with EyeLink 1000 eye tracker (SR-Research, Ottawa, Canada), which samples the gaze position at 1000 Hz using a dark-pupil cornea reflection. The eye tracker was located inside the magnetically shielded room 70 cm from the MEG sensor helmet and below the line of sight of the subject. The eye tracker has been verified not to cause significant artifacts in the MEG signals. Before each experiment, the eye tracker was calibrated using

9 points on the stimulus screen. Fixations and saccades were detected from the raw data, using the method by Stampe (1993) embedded to the EyeLink software. The threshold levels were 30°/s for saccade velocity, 8000°/s^2 for saccade acceleration, and 0.1° for saccade motion.

2.3. Magnetoencephalography (MEG)

MEG signals were measured with a 306-channel neuromagnetometer (Elekta-Neuromag Vectorview; Elekta Oy, Helsinki, Finland) in which 102 sensor units are arranged in a helmet-shaped array and each unit includes two orthogonal planar gradiometers and one magnetometer. The device was located in a three-layer magnetically shielded room (Imedco AG, Hägendorf, Switzerland) at the Brain Research Unit of the O.V. Lounasmaa Laboratory, Aalto University, Espoo, Finland (MEG Core, Aalto NeuroImaging). During the recording, the eye tracker sent trigger signals to the MEG when the gaze entered or left the stimulus image. To suppress magnetic interference, we applied temporally-extended signal space separation (tSSS) with a correlation window of 16 s and a correlation limit of 0.90 (Taulu and Simola, 2006).

2.4. Stimulus presentation

The stimuli were natural images selected from a set used in a previous experiment (Hussain et al., 2014). We scaled the images so that the maximum height was 150 pixels and width 215 pixels, corresponding to 9.4 cm and 13.4 cm, respectively, on the screen (viewing angles 4.3° and 6.1°). The images were grouped into 5 semantic categories as in the previous study (Hussain et al., 2014), each group representing one of the 5 tasks in the experiment. Also images non-relevant for the tasks were included within these groups. Similarly to previous works (Klami et al., 2008; Hardoon and Pasupa, 2010), images were presented to the subjects in a rectangular grid to which both relevant and non-relevant images were randomly placed. In each of the five tasks (see Table 1), 20 collages, each with 16 images placed on a 4x4 grid, were shown to the subject. The relevant images were equally distributed within the grid across the trials.

The task and the collage ordering were shuffled using the Latin-squares rule across the subjects.

The stimulus images were projected on a screen located 130 cm in front of the subject. The stimuli, MEG signals, and gaze signals were synchronized by trigger signals sent from a stimulus computer. Each experiment lasted on average 26 min, excluding instructions, preparations and eye-tracker calibration.

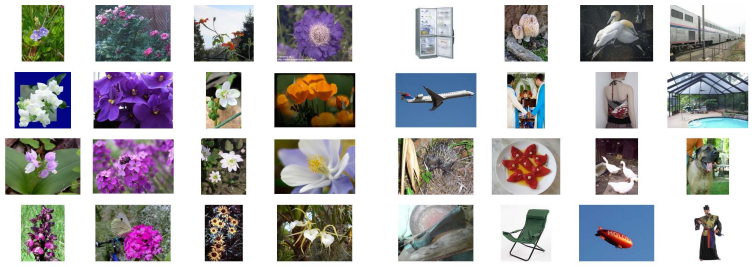
Table 1: Search tasks assigned to the subjects.

	Tasks
1	Select the most attractive sunflower.
2	Find one image most related to man-made flying.
3	Find a cheetah.
4	Select the most attractive deer.
5	Find one image most related to American football.

Figure 1 shows an example collage from each search task. The collages were designed so that a subset of images in each collage was relevant for the given task, whereas the rest were irrelevant. Hence, the ground-truth relevances of the images are known. During the experiment the user was asked to select the most relevant image in each collage by gazing at that image and by clicking a button (by lifting the index finger of the right hand). Then a new collage was displayed (repeated until the end of the task).

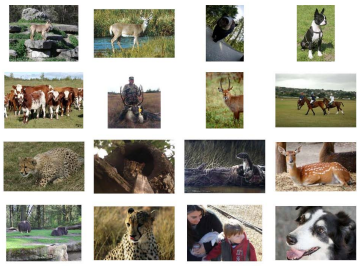
When learning our classifier, we used the ground-truth relevances, available for all images, as the training and test labels. We additionally collected from the users data on which image they considered the most relevant in each collage, in order to keep the user focused on the search task, but this information was not used in our main analyses. In the analysis, we excluded the last 400 ms before the click, to guarantee that decoding is done based on the neural correlates of relevance determination instead of the artifacts caused by the action of clicking itself.

Figure 2 shows an example collage and gaze pattern for the task “find chee-

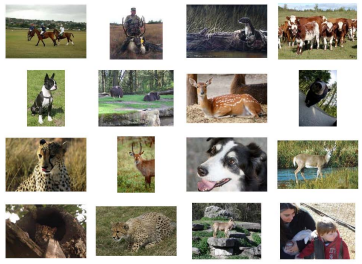


(a) Task 1

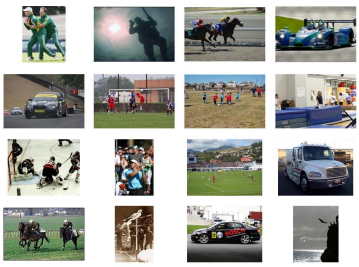
(b) Task 2



(c) Task 3



(d) Task 4



(e) Task 5

Figure 1: Example collages from search tasks listed in Table 1.

tah" for one subject. Although the task of the subject was to find the most relevant image of cheetah (subjective relevance), we used ground-truth labels (objective relevance) to train the classifier. It was thus possible to give the IIR system positive feedback also for images the user was likely to consider as potentially relevant (here the other cheetah image), without making the feed-

back collection procedure too tedious for the user. Importantly, the collage also comprised images that were not fixated at all. These unfixated images and images attracting less than 200 ms gaze time were automatically discarded from our analysis. In a practical IIR system these images could be automatically labeled as "non-relevant" to improve classification performance. See Section 3.1 for more details on how short-time epochs of gaze and MEG data were created based on the gaze pattern.

Figure 2: An example collage and gaze pattern for the task "find cheetah." Fixations and saccades during visual exploration are denoted by red arrows, and the fixations are numbered according to their temporal order. The last fixation point, where the subject pressed a button to select the most relevant image, is denoted by a blue circle. In this example collage, 2 out of 16 images were relevant (i.e., they contained an image of a cheetah) and they are denoted by green rectangles around the images in this illustration (the rectangles were not shown during the experiment). The other 14 images were non-relevant.

The degree of difficulty varied somewhat between the tasks and collages as is obvious from the task definitions and the example collages. To validate that

users mostly chose images from the category "relevant", we computed success rates for each subject. On average, users chose a relevant image with an accuracy of 87.7 %. Although the subjects performed well in most cases, some of the tasks were very challenging. Despite the difficulty of some of the tasks, we expected that the decoding of the ground-truth relevances would still be possible from MEG data for most images.

3. Methods

Our goal in this study was to quantify the feasibility of efficient relevance decoding on the basis of both MEG and gaze data. Because gaze signal based predictors have been succesfully built before, our interest was to construct two particular decoders: i) a decoder based on MEG signal only, and ii) a decoder based on a combination of gaze and MEG signals. The success of the former decoder suggests that image retrieval systems operating with brain signal based relevance feedback is realizable. On the other hand, if the latter decoder turns out to be better than the gaze-based decoder, this suggests that gaze-based image retrieval systems could be improved by utilizing additional information from brain activity.

3.1. Preprocessing

For building the MEG-based classifier, we used the measured gaze signals to identify *epochs*, time intervals when subjects looked at one image. We then extracted the MEG signal of the 200—4870 ms epochs, and discarded epochs corresponding to very short looking times (less than 200 ms) from further analysis, due to their poor spectral resolution. Since information about the relevance was available for each image, we unambiguously associated each epoch either with the "relevant" or the "non-relevant" category. For instance, if the gaze was first targeted inside the borders of a relevant image for 400 ms, then inside the borders of another relevant image for 600 ms, and after this inside the borders of a non-relevant image for 500 ms, three epochs of lengths 400 ms, 600 ms,

and 500 ms were created with the category labels "relevant," "relevant," and "non-relevant". If a subject looked at the same image multiple times, only the epoch corresponding to the longest gaze time for that image was used in the analysis. Note that during visual search, the gaze was often only for a short time over a single image, implying that many of the epochs were relatively short (the mean epoch length was 510 ms). The total number of epochs per subject varied between 431 and 1357 (the average number of epochs across subjects was 905.1). Depending on the subject, the number of "relevant" epochs was 16–23 % of the number of "non-relevant" epochs, i.e., the two category distributions were highly unbalanced for all the subjects. For building the gaze-based classifier, we extracted features from the gaze signal for exactly the same time intervals as for the MEG-based classifier.

3.2. Feature extraction

The features extracted for classification from the epochs of the gaze and MEG signals are explained next.

3.2.1. MEG signal features

It is plausible that visuospatial attention plays a central role in the discrimination between relevant and non-relevant images in the brain. Therefore, we expected that the brain signatures of visuospatial attention are useful in this task. Oscillatory activity from several frequency bands has been previously associated with selective attention (Womelsdorf and Fries, 2007). Therefore, we extracted power features across a wide frequency band and aimed at finding class-discriminative information in them in a data-driven fashion. First, to obtain band-limited MEG signals, we used second-order Butterworth filters with the following cut-off frequencies:

- less than 4 Hz (delta)
- 4–8 Hz (theta)
- 8–12 Hz (alpha)

- 12–30 Hz (beta)
- 30–50 Hz (gamma).

We applied filters to each epoch separately. This epoch-based filtering was preferred over filtering the entire MEG time series at once to avoid introducing artifical temporal dependencies between adjacent epochs; such dependencies would not allow unambiguous labelling of the epochs to relevant vs. non-relevant. After filtering, we computed power features from the filtered epochs (see details below).

A naïve way to use information from multiple frequency bands in the relevance prediction would be to concatenate the power features from indivual frequency bands together. However, this would result in a large number of features and overfitting would be a major concern because of the relatively small number of epochs in each category. To avoid overfitting, we reduced the dimensionality of the feature space to one feature per channel. We did this by applying principal component analysis (PCA) separately for each channel, similarly to an earlier study (Kauppi et al., 2013). This procedure is quite different from the standard approach in machine learning where PCA is used only once to reduce the dimensionality of the original feature space. The three-way structure of the MEG signals (channels $\times$ frequency $\times$ epochs) makes our approach natural, because now PCA captures spectral information which is specific to each channel.

Altogether 48 features were extracted, corresponding to the total number of gradiometer channels in the posterior cortex. Because the brain networks supporting voluntary and stimulus-triggered attention are widely spread (Kastner and Ungerleider, 2000), it is possible that also other brain areas than the parieto–occipital cortex would include discriminative MEG information in this task. However, due to sensitivity of the frontal MEG channels to eye artifacts, we decided to include only parieto–occipital sensors in our analysis.

The details of our MEG signal feature extraction scheme, called "Spectral PCA", are as follows:

- **Estimation of band power features.** Let us denote band-pass filtered epochs as $\mathbf{t}_{c,f}(n)$, for $n = 1, 2, \ldots, N$, where N is the number of epochs, c is the channel index and f is the frequency band index. For each epoch, channel and frequency band, the signal power is computed from the elements of $\mathbf{t}$ as:

$$x_{c,f}(n) = \sum_{i=1}^{T_n} \left(t_{i,c,f}(n)\right)^2, \tag{1}$$

where T_n is the number of time points in the nth epoch. Spectra consisting of the five power features for each channel are denoted by $\mathbf{x}_c(n) \subset \mathbb{R}^5$, for $c = 1, 2, ..., C$, where C is the total number of channels.

- **Transformation of the features.** The power features in each channel were transformed to a logarithmic scale and subsequently standardized to zero mean and unit variance. This way, features in different frequency bands were put into the same scale. The transformed spectra are denoted by $\tilde{\mathbf{x}}_c(n)$.

- **Channel-wise PCA.** We applied PCA channel-wise for the five-dimensional data sets $\tilde{\mathbf{x}}_c(n)$, for $n = 1, 2, \ldots, N$. Note that since we use standardization, PCA finds eigenvectors of the correlation matrices of the above data sets. We denote eigenvectors explaining the highest amount of variance in each channel as $\mathbf{f}_c$, for $c = 1, 2, \ldots, C$.

- **Computation of the final features.**

  We projected each spectrum channel-wise onto the direction of the eigenvectors $\mathbf{f}_c$ to obtain the final features. Thus, the final features are given by the projections: $\mathbf{y}_c(n) = \mathbf{f}_c^{\mathrm{T}} \mathbf{x}_c(n)$, for $c = 1, 2, \ldots, C$. The final features consist of spectral information from the distinct frequency bands.

3.2.2. Gaze signal features

Efficient sets of gaze-signal-based features for relevance prediction have been proposed earlier. In this study, we used 15 of the 19 gaze signal features proposed

by Hussain et al. (2014), leaving out some features not applicable for our setup. The gaze tracker provides raw coordinate measurements at a constant sampling frequency of 1000 Hz and the fixations are extracted from these measurements. Typical gaze analysis operates at the level of fixations, but here we consider also features extracted directly from the raw coordinates (marked as "raw" in Table 2) to improve robustness for potential issues in fixation detection. Table 2 summarizes the used features.

Table 2: Gaze signal features (taken from Hussain et al. (2014)). These features were extracted for each "looked-at" image in each trial.

	Type	**Description**
1	Raw	log of total time of viewing the image
2	Raw	total time for coordinates outside fixations
3	Raw	percentage of coordinates inside/outside fixations
4	Raw	average distance between two consecutive coordinates
5	Raw	number of subimages covered by measurements
6	Raw	maximum pupil diameter during viewing
7	Fixation	total number of fixations
8	Fixation	mean length of fixations
9	Fixation	total length of fixations
10	Fixation	percentage of time spent in fixations
11	Fixation	number of re-visits (regressions) to the image
12	Fixation	length of the first fixation
13	Fixation	number of fixations during the first visit
14	Fixation	distance to the fixation before the first visit
15	Fixation	duration of the fixation before the first visit

3.3. Image relevance decoding by Bayesian classifiers

We used Bayesian classifiers to decode image relevance. The probabilistic nature of Bayesian classifiers brings us certain benefits. In particular, they provide an uncertainty measure for their predictions, unlike other common linear

models such as an LDA and non-linear models such as the support vector machine (SVM). Probabilistic estimates for the relevance are valuable for retrieval tasks, since they allow ranking images according to the relevance. An additional benefit of having probabilistic predictions is that "late fusion" (Ye et al., 2012) of multiple classifiers is possible in the probabilistic framework. In the late fusion, two or more classifiers are trained using different feature sets. Then, in the classification stage, the class probabilities of unseen test samples are computed separately for each classifier, and the average of these class probabilities is adopted as the final class probability.[2]

The common trend in the neuroimaging community is to use linear classifiers for two reasons: i) Brain imaging data sets are prone to overfitting due to their drastically low signal-to-noise ratio, which can be handled by simple linear models that have a very strong bias. ii) In linear models the learned regressor weights reveal which features contribute to the classification. However, we will show in Section 4.1 that it is possible to also learn interpretable *non-linear* models from highly noisy MEG data without overfitting, and thus be as interpretable as linear models with an additional boost in classification performance. We used Gaussian processes (GP) (Rasmussen and Williams, 2005) for classification; GPs give probabilistic and non-parametric models providing uncertainty predictions. It has been previously suggested that the GP classifier is well-suited to analyze neuroimaging data sets, because it automatically finds a trade-off between the data fit and regularization (Zhong et al., 2008; Boyu et al., 2009). In this way, the method can adapt even for high-dimensional and non-linear brain imaging data without overfitting. Because the idea of using the GP in the field of neuroimaging is not widespread, we introduce details of the GP classifier in the Appendix.

We trained the GP classifier with the radial basis function (RBF) kernel with diagonal covariance, also known as the automatic relevance determination

[2]Taking the average here simply means assigning a uniform prior over the classifiers at hand.

(ARD) kernel (see Appendix). The benefit of the ARD kernel is that it enables investigation of the contribution of features for predictions. We used late fusion of two GP classifiers, one trained based on the MEG signal features, and the other based on the gaze signal features, and hypothesized that the late fusion of these feature sets improves the classification performance. Another possibility for combining the MEG and gaze classifiers would be to concatenate the corresponding feature sets (early fusion), but we observed this brought suboptimal performance. We compared the performance of our GP classifier to the Bayesian linear logistic regression (LR) model (see Appendix for details).

3.4. Performance estimation and significance testing

Our search paradigm, which involves continuous exploration of a visual scene, requires careful design of both performance estimation of the classifier and assessment of the results. One difficulty is that the sizes of the two categories are highly unbalanced, because several non-relevant images often need to be browsed before a relevant image is found, implying that most epochs in each collage are labelled as "non-relevant." Because of this, we used performance metrics immune to the class balance. We evaluated the performance of the classifiers using two metrics (Fawcett, 2006):

- **AUC-ROC**: Area Under Receiver Operating Characteristics (ROC) Curve.
- **AUC-PR:** Area Under Precision-Recall (PR) Curve.

The ROC curve draws the change of *recall*[3] $\left(\frac{TP}{TP+FN}\right)$ as a function of *fall-out* $\left(\frac{FP}{FP+TN}\right)$ for varying decision thresholds. The PR curve draws the change in *precision* $\left(\frac{TP}{TP+FP}\right)$ as a function of recall for varying decision thresholds. The areas under these curves measure the discriminative power of a binary classifier independently of the decision thresholds. These metrics are

[3] In the definitions of *precision*, *recall* and *fall-out*, we use the following shorthand notations: *TP=true positives*, *FP=false positives*, *FN = false negatives*, and *TN=true negatives*.

well-suited to our imbalanced data set, since they consider all possible decision thresholds. While AUC-ROC is a widely accepted performance measure in binary classification, AUC-PR may be a better alternative for data sets with highly unbalanced category distributions (Sonnenburg et al., 2007). Classification accuracy, even though being widely used in brain decoding studies, would not be the most appropriate measure here since it would strongly favor majority voting.

For AUC-ROC, the perfect classifier would have a score of 1 and a random classifier a score of 0.5. For AUC-PR, the limits depend on the class ratio and hence vary from subject to subject. Values clearly above the baseline indicate that the classifier was able to differentiate between relevant and non-relevant images within the set of images that were studied sufficiently long (at least for 200 ms); for a practical IIR system, the accuracy would be higher since the images viewed only briefly could be labeled as non-relevant.

When estimating classifier performance, it is highly important to ensure that training and test data sets are independent to avoid bias. Because of the continuous nature of the task, adjacent epochs may share partially same information. However, epochs extracted from different collages hardly have any dependencies due to the break between the presentation of the collages. For this reason, we assessed the performance using a block-wise four-fold cross-validation (CV) scheme, where the epochs of the training data were always drawn from different collages than those of the test data.

Because our decoding task is expected to be very challenging due to the low signal-to-noise ratio of unaveraged MEG data, it is of prime importance to use a proper testing scheme to assess the significance of the findings. Recently, it was shown with simulated brain imaging data that existing parametric tests can easily indicate significant performance even though there is no discriminative information in the data (Schreiber and Krekelberg, 2013). Instead, the recommended test is a nonparametric permutation test under the null hypothesis that there is no association between epochs and categories (Schreiber and Krekelberg, 2013). We approximated a distribution of the expected classifica-

tion performance under this null hypothesis by randomly shuffling the category labels of the training data before training and testing the classifier. The procedure was repeated 200 times. We used a block-wise shuffling to preserve the dependency structure of the original data also in the shuffled data, as suggested by Schreiber and Krekelberg (2013). The permutation distribution and subsequent significance testing was performed for both of the performance metrics.

4. Results

4.1. Classification performance

Table 3 shows the performance for the GP classifier (RBF kernel with diagonal covariance matrix) based on the three evaluated feature sets: "MEG"=MEG signal features, "gaze"=gaze signal features, and "combined"=mean of the predictions of the above two classifiers (late fusion). We report average and subject-wise results for both metrics, AUC-ROC and AUC-PR. The mean AUC-ROC across the subjects was 0.654 for the "MEG," 0.655 for the "gaze," and 0.679 for the "combined." The corresponding mean AUC-PR values for these feature sets were 0.348, 0.363, and 0.386. The entire precision-recall curves of the subjects for the classifier "combined" are provided as supplementary material (see Fig. S1). For AUC-ROC the "MEG" feature set improved on chance level for 5 users, the "gaze" set for 6 users, and the late fusion result for 7 users. The corresponding numbers for the AUC-PR metric were: 5, 7, and 8. The combined classifier improved the AUC-ROC performance significantly compared to the gaze signal feature set alone (paired t-test; $p = 0.0040$). The corresponding improvement was also significant based on the AUC-PR metric (paired t-test; $p = 0.0107$).

To validate the choice of our nonlinear GP classifier, we compared its performance against the Bayesian linear LR classifier. The GP classifier was significantly better for both metrics (mean AUC-ROC across the subjects 0.654 for GP vs. 0.609 for LR, paired t-test $p = 0.0013$; mean AUC-PR across the subjects 0.348 vs. 0.270, paired t-test; $p = 0.0001$).

Table 3: Performance of the GP classifier for each subject based on two performance metrics: AUC-ROC = area under ROC curve, and AUC-PR = area under precision-recall curve. The results, which are significantly above a random chance level (block permutation test; $p < 0.05$; Bonferroni corrected), are denoted by an asterisk. The feature sets are named as follows: MEG = MEG signal feature set, gaze = gaze signal feature set, combined = mean of the predictions of the gaze and MEG classifiers (late fusion). For AUC-ROC, a random baseline is 0.50 for each subject. For AUC-PR, the baseline corresponds to a class ratio. Because the baseline values of the AUC-PR are different for each subject, they are reported in a separate column of the table.

	AUC-ROC			**AUC-PR**			
	MEG	gaze	combined	MEG	gaze	combined	baseline
S1	0.589	0.561	0.601*	0.334*	0.293*	0.353*	0.229
S2	0.678*	0.659*	0.691*	0.390*	0.378*	0.417*	0.154
S3	0.677*	0.683*	0.710*	0.399*	0.363*	0.403*	0.226
S4	0.681	0.609	0.673	0.230	0.198	0.241	0.165
S5	0.602	0.622*	0.647*	0.341	0.345*	0.366*	0.159
S6	0.758*	0.766*	0.781*	0.460*	0.502*	0.508*	0.218
S7	0.706*	0.757*	0.760*	0.373	0.450*	0.454*	0.227
S8	0.544	0.536	0.538	0.231	0.266	0.255*	0.167
S9	0.654*	0.704*	0.710*	0.379*	0.475*	0.477*	0.227
avg	**0.654**	**0.655**	**0.679**	**0.348**	**0.363**	**0.386**	**0.197**

4.2. Contribution of the MEG channels to prediction

Figure 3 shows the spatial distribution of the contribution of the MEG channels to prediction, measured by the learned precisions ($1/\sigma_d^2$) for 6 subjects. Subjects 4, 5 and 8 were excluded from this analysis because the classifiers of these subjects did not yield reliable results (see Table 3). The feature precisions provide information about which features were the most capable of discriminating between the categories. However, it should be noted that most discriminative features do not directly imply high neural activity (Haufe et al., 2014). The number and location of the features showing high precision values were relatively variable across subjects, but there were also similarities; for instance,

subjects 1, 2 and 3 showed high feature precision in the left lateral occipital cortex.

In this study, we chose the RBF kernel with a diagonal covariance matrix to enable visualization of the feature precisions over the MEG helmet as shown in Fig. 3. However, an isotropic covariance matrix would allow faster fitting of the model and could therefore be more useful in real-time applications. To investigate whether the isotropic covariance matrix would yield comparable performance to the diagonal one, we trained the GP classifier also using the isotropic covariance matrix. In this case, the mean AUC-ROC and AUC-PR performance was 0.658 and 0.352, respectively. These values were slightly higher than the ones obtained with the diagonal covariance (0.654 for the AUC-ROC and 0.348 for the AUC-PR), but the performances with the two covariance matrices were not significantly different (paired t-test; $p = 0.613$ for the AUC-ROC and $p = 0.543$ for the AUC-PR).

4.3. Validation of the MEG features

We validated our feature extraction scheme by evaluating the classification performance also on the basis of other MEG feature sets. Also these feature sets were extracted from the occipital gradiometer channels. The seven feature sets included: the total power of the MEG signal (denoted as "total"), band powers from five distinct frequency bands ("delta," "theta," "alpha," "beta," and "gamma") and the concatenation of the individual band powers ("all"). Figure 4 shows the classification performance computed on the basis of these feature sets together with the proposed "PCA" feature set. It can be seen that the proposed feature set provided the highest mean performance in terms of both AUC-ROC (Fig. 4(A)) and AUC-PR (Fig. 4(B)). The difference is significant for 5 out of the 7 feature sets (Table 4).

5. Discussion

We analyzed MEG signals and eye gaze from nine subjects who explored collages of natural scene images. Only a subset of the images in each collage

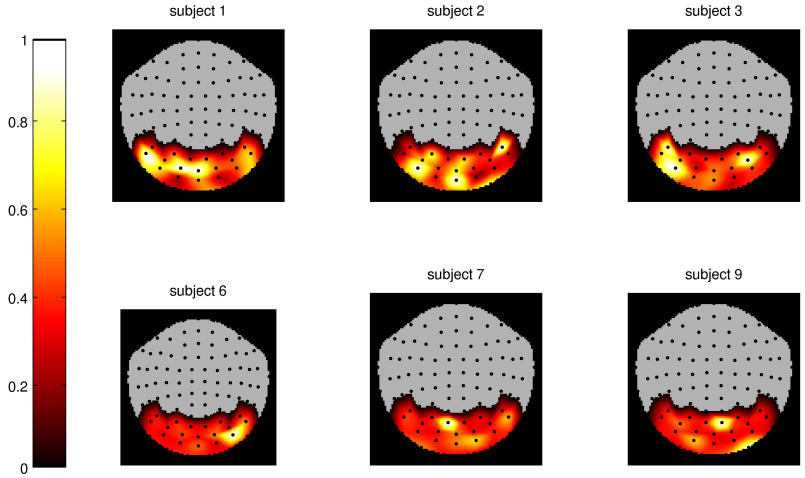


Figure 3: Contributions of the brain regions to GP classification, measured by the learned feature precisions of the RBF kernel with diagonal covariance. Bright yellow corresponds to the highest and dark red to the lowest importances found. The values are reported as means across folds. The orientation of the helmets is as follows: top of the array = anterior parts of the brain, bottom = posterior, right on right. Each spatial location in the helmet contains a pair of gradiometer sensors; only the sensor with the higher weight is shown for each pair. MEG signals from the gray area were not used in the classification in order to restrict the analysis to the occipital cortex, and hence to avoid classification based on eye artifacts that may be present at the frontal sensors.

Table 4: Comparison of the classification performance of the proposed feature extraction scheme (denoted as "PCA" in Fig. 4) against the other MEG signal feature sets. The results with "PCA" were compared against results of the seven feature sets using the paired t-test. The table contains the original p-values, and significant differences compared to the "PCA" after the Bonferroni correction have been marked by an asterisk.

	total	**delta**	**theta**	**alpha**	**beta**	**gamma**	**all**
AUC-ROC	0.0005*	0.0003*	0.0012*	0.0009*	0.0632	0.0341	0.0002*
AUC-PR	0.0001*	0.0002*	0.0029*	0.0001*	0.0612	0.1480	0.0047*

was relevant to the given task, and the subjects were asked to pick the most relevant image by directing their gaze to it and simultaneously clicking a button.

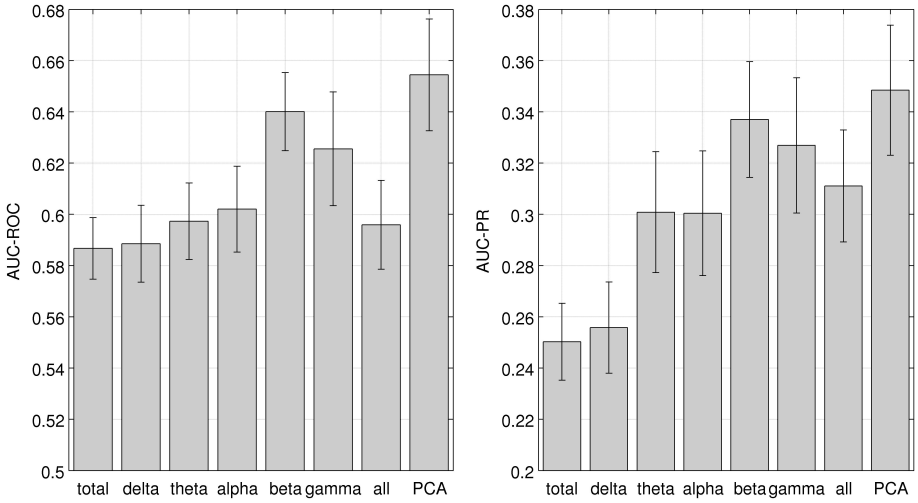


Figure 4: Classification performance of eight different MEG feature sets based on A) AUC-ROC = area under ROC curve, and B) AUC-PR = area under precision-recall curve. From left to right, bars correspond to the results (mean +/- standard error across subjects) obtained by training the GP classifier with the following eight MEG feature sets: total = total signal power, delta = delta-band power (0–4 Hz), theta = theta-band power (4–8 Hz), alpha = alpha-band power (8–12 Hz), beta = beta-band power (12–30 Hz), gamma = gamma-band power (30–50 Hz), all = a concatenation of the five band-power feature sets, and PCA = Spectral PCA (see Section 3.2.1).

Our goal was to analyze whether it would be possible to decode relevance of the image by utilizing only information about brain activity. To this end, we extracted short MEG epochs from the time intervals when the subjects looked at the images and attempted to classify the epochs as "relevant" or "non-relevant" using a GP classifier.

Our study showed that the relevance of an image the subject was looking at could be predicted from the MEG signal alone significantly better than chance for most of the subjects (see Table 3). Moreover, when combining the MEG and gaze signals using late fusion, the classification performance is improved compared to using either of these two modalities alone. These results imply that it might be possible to build implicit relevance feedback mechanisms for image

retrieval systems which are based on integrated information of gaze and brain activity, or even brain-activity alone, thereby considerably extending previous gaze-based studies in information retrieval (Puolamäki et al., 2005; Salojärvi et al., 2005; Hardoon and Pasupa, 2010; Ajanki et al., 2009).

In this study we used objective labels (based on the ground truth given the task) for learning the relevance classifier instead of using subjective labels based on user input. We made this choice because it is very difficult to collect reliable subjective feedback on all images from subjects in a natural manner. The objective labels are a good basis for learning the relevance estimator since they approximate well the set of images about which the user needs to make a conscious relevance decision, potentially even more so than the subjective relevance estimates would. The eventual estimator will still be applicable for IR tasks without ground-truth labels and it will provide useful feedback for the retrieval system.

In neuroimaging reseach, linear classifiers, such as a linear SVM and LDA are often favored over non-linear ones, because the simplicity of linear classifiers helps to prevent overfitting on the usually very noisy and limited neuroimaging data (Mur et al., 2009; Müller et al., 2003). However, also non-linear classifiers have been successfully utilized in BCI (Müller et al., 2003) and neuroscientific research (Davatzikos et al., 2005; Cox and Savoy, 2003), suggesting that classification problems of complex neuroimaging data sets may be intrinsically non-linear. If this is the case, non-linear classifiers should outperform linear ones as long as a classifier is properly regularized against overfitting. For many popular non-linear classifiers, such as the SVM with the RBF kernel, efficient model regularization is difficult due to the extensive hyperparameter space. However, our results suggest that the GP classifier is capable of learning relevant non-linearities in data while tuning the kernel hyperparameters without requiring extensive cross-validation and without overfitting. Moreover, as shown in Fig. 4, the GP classifier enables the investigation of the importances of the features in a similar manner as classification coefficients in linear models. This is a great benefit in neuroimaging studies, in which interpretation of the classifiers has

primary importance.

The use of a free-viewing search paradigm, and a multivariate classification approach for the analysis of the relevance of complex natural images, may also open new possibilites to study brain mechanisms related to stimulus relevance. Traditional univariate approaches, such as the analysis of power differences in single MEG channels, would be rather uninformative in a naturalistic setup such as ours, where the subject's behavior is highly uncontrolled and leads to a large variation in strategies and responses across trials and subjects. Because our relevance decoder takes simultaneously into account complex statistical properties of multiple sensors, it can reveal discriminative features even when they are not significant at the level of single MEG sensors.

In the visual image exploration task used here, a subject selected relevant images based on both visual properties of those images as well as their expectations and knowledge about the images. Hence, it is likely that both top-down and bottom-up attentional control mechanisms (Egeth and Yantis, 1997) were in use during the search task, involving several brain areas simultaneously. However, we used for classification only the posterior MEG channels to avoid contamination by eye artifacts. This selection was important in our study because we wanted to investigate the information from eye movements and brain activity separately.

As a feature extraction scheme, we proposed projecting the powers of multiple frequency bands onto the direction of the first principal axes using PCA, separately for each MEG channel. Our method provided the highest classification performance among the other tested MEG signal feature sets. Interestingly, our scheme provided significantly better results when compared with the classifier trained using all the band-power features (see Table 4), showing that the method could integrate discriminative information from a wide frequency band by simultaneously preventing overfitting. Besides high performance, an additional benefit of using the Spectral PCA feature extraction is that that the frequency band of interest does not need to be strictly specified *a priori*, making it an applicable tool for different decoding tasks.

Also the results of "beta" and "gamma" band-power feature sets provided good results in our tests (see Table 4). Active role of these frequency bands in attentional mechanisms has been suggested earlier (Womelsdorf and Fries, 2007). The total power of the MEG signals provided the lowest classification performance, emphasizing the importance of rhythmic activity measured from distinct frequency bands in our task. Also the results of the lowest frequency band were relatively poor. This was not surprising given the short length of many filtered epochs.

In this study, we concentrated on features of rhythmic brain activity present in the frequency domain of the MEG signals. We favored spectral features because they are easy to compute for epochs of varying lengths and because they have been successfully used to control attention-based BCIs (van Gerven and Jensen, 2009). In fact, the detection of some key temporal features, such as the P300 response (Blankertz et al., 2011), would not have been possible for many epochs due to their short duration. Nevertheless, it is possible that temporal features of the MEG signals would provide additional discriminative information. However, due to the highly complex spatiotemporal structure of the epochs, an entire new study with a different design would be needed to examine the effect of the most appropriate temporal features on the classification performance.

In summary, we have successfully demonstrated that it is possible to decode relevance information of natural images from MEG signals during unrestricted visual search using multivariate non-linear modeling, and that the integration of MEG and gaze signal information allows more powerful image retrieval when compared with gaze signals alone. We expect that these findings will be highly relevant in the future development of brain-activity-based information retrieval systems.

Appendix

Here, we present the relevant theory of the Bayesian relevance predictors used in this paper. We choose the Bayesian framework due to its well-known benefits: i) Averaging over a learned distribution of model parameters rather than learning a point estimate, ii) avoiding overfitting, and iii) making predictions together with an uncertainty score.

Gaussian process classifier

A Gaussian process $\mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is a stochastic process determined by a mean function $\mu(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$, both defined on the entire feature space. A Gaussian process serves as a prior over the space of functions $f(\mathbf{x})$ that map the feature space to the output space. A Gaussian process prior on output points $\mathbf{y} = \{y_1, \cdots, y_N\}$ of any chosen finite set of observations $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ is $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K})$, where $\boldsymbol{\mu} = \{\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_N)\}$ and $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ for a kernel function $k(\cdot, \cdot|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. For a given set of input and real-valued output observations $\mathbf{X}$ and $\mathbf{y}$, the predictive distribution is analytically available as

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(y_*|\mathbf{k}_*^T\mathbf{K}^{-1}\mathbf{y}, k_* - \mathbf{k}_*^T\mathbf{K}^{-1}\mathbf{k}_*),$$

where $\{\mathbf{x}_*, y_*\}$ is a newly seen observation and its corresponding target, $\mathbf{k}_*$ is the vector of kernel responses between $\mathbf{x}_*$ and each training point, and k_* is the kernel response of $\mathbf{x}_*$ to itself. For categorical output, as for binary classification, a Gaussian process prior is applied on the mapping function from the input observations to the latent decision margin variables $\mathbf{f} = \{f_1, \cdots, f_N\}$, which are then converted to probabilities by being squeezed by a sigmoid function:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}),$$
$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} \sigma(y_i f_i),$$

where $\sigma(z) = 1/(1 + \exp(-z))$. Since the sigmoid function is not conjugate with the normal distribution, the predictive distribution for this model is not

analytically tractable. Hence, approximation techniques are required for inference. We choose to approximate the posterior by the Laplace's method due to its computational efficiency: $p(\mathbf{f}|\mathbf{X},\mathbf{y}) \approx \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{H}_{\hat{\mathbf{f}}})$, where $\hat{\mathbf{f}}$ is the posterior mode and $\mathbf{H}_{\hat{\mathbf{f}}}$ is the Hessian at the posterior mode. The posterior mode can easily be calculated by taking the gradient of the log-posterior and using any gradient-based optimizer.

The Laplace approximation corresponds to a second-order Taylor expansion as shown by Rasmussen and Williams (2005). We chose the second-order expansion because the normal density is a quadratic function of the random variable of interest. The first-order term tends to zero since the expansion is evaluated at the mode. The resulting formula approximates the posterior distribution with a multivariate normal distribution with full covariance. Higher-order expansions would technically be possible but would bring a considerable computational burden.

An elegant property of the GP is that its probabilistic nature allows for a principled way of fitting kernel hyperparameters $\boldsymbol{\theta}$ to data, unlike non-probabilistic kernel-based learners such as support vector machines (SVMs) that can only fit those parameters by grid search over a validation set. The kernel hyperparameters can be fit by maximizing the log-marginal likelihood $\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, which is the well-known Type II maximum likelihood technique. For GP classification, the intractable log-marginal likelihood can be approximated by a Taylor expansion as follows:

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \simeq p(\hat{\mathbf{y}},\hat{\mathbf{f}}|\mathbf{X},\boldsymbol{\theta}) \int \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{f}-\hat{\mathbf{f}})^{\mathbf{T}}\mathbf{H}_{\hat{\mathbf{f}}}(\mathbf{f}-\hat{\mathbf{f}})\right) \mathbf{df}.$$

Here we have a Gaussian integral, which is analytically tractable. As a result, we have

$$\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \simeq -\frac{1}{2}\hat{\mathbf{f}}^T\mathbf{K}^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) + \frac{1}{2}\log|\mathbf{K}\mathbf{H}_{\hat{\mathbf{f}}}|.$$

The gradient of this function with respect to $\boldsymbol{\theta}$ is also analytically tractable. Hence, the log-posterior and kernel hyperparameters can easily be learned jointly by a gradient-search in a coordinate-ascent fashion (i.e. iterating between keep-

Table 5: Free parameters of our Gaussian process classifier and the related methods used to fit them to data are listed.

Parameter	Description	Learning Method
$\boldsymbol{\mu}(\mathbf{x}) = C$	Constant mean function	Gradient descent
$\boldsymbol{\theta} = [\sigma_1^2, \cdots, \sigma_D^2]$	Kernel hyperparameters (see Sec. 5)	Gradient descent
$\mathbf{f} = f_1, \cdots, f_N$	Latent decision margin variables	Laplace approximation

ing one fixed and learning the other). We also choose a constant mean function $\mu(\mathbf{x}) = C$, and fit also C using Type II maximum likelihood, simply by taking another derivative of the log-marginal likelihood with respect to C. The free parameters of the GP and the methods used for learning them are summarized in Table 5 More details on how to solve machine learning problems using GPs are available in Rasmussen and Williams (2005).

Given a trained GP classifier, a new observation $\mathbf{x}_*$ can be classified by calculating the class-conditional distribution $p(y_* = +1|\mathbf{x}_*) = \mathbb{E}[\mathbf{f}_*|\mathbf{y}, \mathbf{X}, \mathbf{X}^*]$. Even though this expectation is not analytically solvable, its mean $\mathbf{k}_*^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ gives a reasonable estimate.

The choice of the kernel function

Assuming that our data consists of Gaussian-distributed chunks, we adopt the radial basis function (RBF) as the kernel function

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\mathbf{x}^T \mathbf{S}^{-1} \mathbf{x}'\right).$$

For the covariance matrix $\mathbf{S}$, we consider two choices:

- Isotropic covariance: $\mathbf{S} = \sigma^2 \mathbf{I}$,
- Diagonal covariance: $\mathbf{S} = diag(\sigma_1^2, \cdots, \sigma_D^2)$

where D is the data dimensionality and the $diag(\cdot)$ constructs a diagonal matrix with the entries given as the arguments. While the isotropic covariance is faster to fit, since it has only one hyperparameter ($\boldsymbol{\theta} = \{\sigma^2\}$), it has the disadvantage of not allowing feature importance analysis. The diagonal covariance, on the

other hand, has a larger hyperparameter set ($\boldsymbol{\theta} = \{\sigma_1^2, \cdots, \sigma_D^2\}$) that consists of one hyperparameter σ_d^2 per data dimension, which is inversely proportional to how large an effect that dimension has in the model fit. This technique is called *automatic relevance determination* (ARD). We use the diagonal covariance to investigate the contributions of MEG channels to *non-linear* prediction of relevance.

Bayesian logistic regression with ARD prior

As an alternative linear classifier, we tested the standard Bayesian logistic regression with ARD prior on regressor weights (Jaakkola and Jordan, 2000). The generative process of the classifier is as follows:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \alpha_d^{-1}) \tag{2}$$

$$p(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_d|a, b) \tag{3}$$

$$p(y_i = +1|\mathbf{w}, \mathbf{x}_i) = \prod_{i=1}^{N} \sigma(\mathbf{x}_i^T \mathbf{w}) \tag{4}$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic function. Equation 2 tells that the regressor weight w_d of each feature dimension d is assigned a normal prior with zero mean and a precision α_d specific to that dimension. This precision is also given a Gamma hyperprior (Equation 3). Hyperparameters of this hyperprior are set for all dimensions to $a = 10^{-2}$ and $b = 10^{-3}$, which leads to $\mathbb{E}[\alpha_d|a, b] = a/b = 10$, imposing an expected prior variance of 0.1. Equations 2 and 3 form the ARD prior, which induces a zero mean and low variance prior over the regressor weights w_d, forcing as many of them to zero as possible. This way, the model both performs feature selection, and automatically adapts its complexity to data. The resulting model turns out to be the probabilistic counterpart of the Lasso regression (Tibshirani, 1996) with the additional benefit of not being heavily dependent on the chosen regularization coefficient.

The inference of the model parameters is performed by calculating the posterior $p(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{X})$. Since this posterior is not available in a closed form, we

approximate it by a factorized distribution $q(\mathbf{w}, \boldsymbol{\alpha}) = q(\mathbf{w})q(\boldsymbol{\alpha})$, following the standard mean field variational approximation scheme. Given the approximate posterior, the output y^* of a newly seen observation $\mathbf{x}$ can be calculated as follows:

$$p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) = \int\int p(y^*|\mathbf{w}, \boldsymbol{\alpha}, \mathbf{x}^*)p(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{X}, \mathbf{y})d\mathbf{w}d\boldsymbol{\alpha}$$
$$\approx p(y^*|\mathbf{w}, \boldsymbol{\alpha}, \mathbf{x}^*)q(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{X}, \mathbf{y})d\mathbf{w}d\boldsymbol{\alpha}.$$

For more details, see Jaakkola and Jordan (2000).

Acknowledgement

Financially supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, grant #251170; and personal grants #131483, #263800, and #266969; grant LASTU #135198), and by the European Research Council (Advanced Grant #232946).

References

Ajanki A, Hardoon DR, Kaski S, Puolamäki K, Shawe-Taylor J. Can eyes reveal interest? Implicit queries from gaze patterns. User Model User-Adap Inter 2009;19(4):307–39.

Arapakis I, Konstas I, Jose JM. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In: Proceedings of the 17th ACM International Conference on Multimedia. ACM; 2009. p. 461–70.

Blankertz B, Lemm S, Treder M, Haufe S, Müller KRR. Single-trial analysis and classification of ERP components–a tutorial. NeuroImage 2011;56(2):814–25.

Boyu W, Feng W, Peng-Un M, Pui-In M, Mang-I V. EEG signals classification for brain computer interfaces based on Gaussian process classifier. In: 7th International Conference on Information, Communications and Signal Processing. ICICS; 2009. p. 1–5.

Buscher G, Dengel A, Biedert R, Elst LV. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. ACM Trans Interact Intell Syst 2012;1(2):9.

Claypool M, Brown D, Le P, Waseda M. Inferring user interest. IEEE Internet Comput 2001a;5(6):32–9.

Claypool M, Le P, Wased M, Brown D. Implicit interest indicators. In: Proceedings of the 6th International Conference on Intelligent User Interfaces. ACM; 2001b. p. 33–40.

Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI)"brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 2003;19(2):261–70.

Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surv 2008;40(2):5.

Davatzikos C, Ruparel K, Fan Y, Shen D, Acharyya M, Loughead J, Gur R, Langleben DD. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. NeuroImage 2005;28(3):663–8.

Dias JC, Sajda P, Dmochowski JP, Parra LC. EEG precursors of detected and missed targets during free-viewing search. J Vision 2013;13(13):13.

Egeth HE, Yantis S. Visual attention: Control, representation, and time course. Annu Rev Psychol 1997;48(1):269–97.

Eugster MJA, Ruotsalo T, Spapé MM, Kosunen I, Barral O, Ravaja N, Jacucci G, Kaski S. Predicting term-relevance from brain signals. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval. 2014. p. 425–34.

Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27(8):861–74.

Fox S, Karnawat K, Mydland M, Dumais S, White T. Evaluating implicit measures to improve web search. ACM T Inform Syst 2005;23(2):147–68.

van Gerven M, Jensen O. Attention modulations of posterior alpha as a control signal for two-dimensional brain–computer interfaces. J Neurosci Meth 2009;179(1):78–84.

Hardoon DR, Pasupa K. Image ranking with implicit feedback from eye movements. In: Proceedings of 2010 Symposium on Eye-Tracking Research & Applications. ACM; 2010. p. 291–8.

Hari R, Salmelin R. Magnetoencephalography: From SQUIDs to neuroscience: Neuroimage 20th anniversary special edition. NeuroImage 2012;61(2):386–96.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 2014;87:96–110.

Haynes J, Rees G. Decoding mental states from brain activity in humans. Nat Rev Neurosci 2006;7(7):523–34.

Huang Y, Liu Q, Zhang S, Metaxas DN. Image retrieval via probabilistic hypergraph ranking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2010. p. 3376–83.

Hussain Z, Klami A, Kujala J, Leung A, Pasupa K, Auer P, Kaski S, Laaksonen J, Shawe-Taylor J. PinView: Implicit feedback in content-based image retrieval 2014;`arXiv:1410.0471`.

Jaakkola TS, Jordan MI. Bayesian parameter estimation via variational methods. Stat Comput 2000;10(1):25–37.

Jangraw DC, Wang J, Lance BJ, Chang SF, Sajda P. Neurally and ocularly informed graph-based models for searching 3D environments. J Neural Eng 2014;11(4):046003.

Joachims T, Granka L, Pan B, Hembrooke H, Gay G. Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2005. p. 154–61.

Kastner S, Ungerleider LG. Mechanisms of visual attention in the human cortex. Annu Rev Neurosci 2000;23(1):315–41.

Kauppi JP, Parkkonen L, Hari R, Hyvärinen A. Decoding magnetoencephalographic rhythmic activity using spectrospatial information. NeuroImage 2013;83:921–36.

Klami A, Saunders C, de Campos TE, Kaski S. Can relevance of images be inferred from eye movements? In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. ACM; 2008. p. 134–40.

Lan T, Yang W, Wang Y, Mori G. Image retrieval with structured object queries using latent ranking SVM. In: Computer Vision–ECCV 2012. Springer; 2012. p. 129–42.

Lew MS, Sebe N, Djeraba C, Jain R. Content-based multimedia information retrieval: State of the art and challenges. ACM T Multim Comput 2006;2(1):1–19.

Moshfeghi Y, Jose JM. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2013. p. 133–42.

Moshfeghi Y, Pinto LR, Pollick FE, Jose JM. Understanding relevance: An fMRI study. In: Advances in Information Retrieval. Springer; 2013. p. 14–25.

Müller KR, Anderson CW, Birch GE. Linear and nonlinear methods for brain-computer interfaces. IEEE Trans Neural Syst Rehabil Eng 2003;11(2):165–9.

Mur M, Bandettini PA, Kriegeskorte N. Revealing representational content with pattern-information fMRI—an introductory guide. Soc Cogn Affect Neurosci 2009;4(1):101–9.

Pohlmeyer EA, Wang J, Jangraw DC, Lou B, Chang SF, Sajda P. Closing the loop in cortically-coupled computer vision: a brain–computer interface for searching image databases. J Neural Eng 2011;8(3):036025.

Puolamäki K, Salojärvi J, Savia E, Simola J, Kaski S. Combining eye movements and collaborative filtering for proactive information retrieval. In: Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2005. p. 146–53.

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. The MIT Press, 2005.

Ruthven I. Interactive information retrieval. Annu Rev Inform Sci 2008;42(1):43–91.

Salojärvi J, Puolamäki K, Simola J, Kovanen L, Kojo I, Kaski S. Inferring relevance from eye movements: Feature extraction. Publications in Computer and Information Science A82; Helsinki University of Technology; Espoo, Finland; 2005.

Saracevic T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. J Assoc Inf Sci Technol 2007;58(13):2126–44.

Schreiber K, Krekelberg B. The statistical analysis of multi-voxel patterns in functional imaging. PLoS ONE 2013;8(7):e69328.

Soleymani M, Chanel G, Kierkels JJ, Pun T. Affective ranking of movie scenes using physiological signals and content analysis. In: Proceedings of the 2nd ACM Workshop on Multimedia Semantics. ACM; 2008. p. 32–9.

Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. BMC Bioinformatics 2007;8(Suppl 10):S7.

Stampe DM. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behav Res Meth Ins C 1993;25(2):137–42.

Taulu S, Simola J. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys Med Biol 2006;51(7):1759.

Tibshirani R. Regression shrinkage and selection via the LASSO. J Royal Stat Soc Series B (Methodological) 1996;58(1):267–88.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain–computer interfaces for communication and control. Clin Neurophysiol 2002;113(6):767–91.

Womelsdorf T, Fries P. The role of neuronal synchronization in selective attention. Curr Opin Neurobiol 2007;17(2):154–60.

Ye G, Liu D, Jhuo IH, Chang SF. Robust late fusion with rank minimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2012. p. 3021–8.

Zhong M, Lotte F, Girolami M, Lecuyer A. Classifying EEG for brain computer interfaces using Gaussian processes. Pattern Recogn Lett 2008;29(3):354–9.