NeuroImage 42 (2008) 1056–1068

Contents lists available at ScienceDirect

NeuroImage

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

Predicting the recognition of natural scenes from single trial MEG recordings of brain activity

Jochem W. Rieger [a,*], Christoph Reichert [a], Karl R. Gegenfurtner [b], Toemme Noesselt [a], Christoph Braun [c], Hans-Jochen Heinze [a,d], Rudolf Kruse [e], Hermann Hinrichs [a]

[a] *Department of Neurology II, Otto-von-Guericke University, Leipziger Str. 44, 39120 Magdeburg, Germany*
[b] *Department of Psychology, Giessen University, Otto-Behaghel-Str. 10, 35394 Giessen, Germany*
[c] *MEG-Center, Eberhard-Karls University, Otfried-Müller-Str. 47, 72076 Tübingen, Germany*
[d] *Department of Behavioral Neurology, Leibniz Institute for Neurobiology, Brenneckestr. 6, 39118 Magdeburg, Germany*
[e] *Department of Knowledge Processing and Language Engineering, Otto-von-Guericke University, Magdeburg, Germany*

ARTICLE INFO

Article history:
Received 5 December 2007
Revised 8 May 2008
Accepted 9 June 2008
Available online 24 June 2008

ABSTRACT

In our daily life we look at many scenes. Some are rapidly forgotten, but others we recognize later. We accurately predicted recognition success with natural scene photographs using single trial magnetoencephalography (MEG) measures of brain activation. Specifically, we demonstrate that MEG responses in the initial 600 ms following the onset of scene photographs allow for prediction accuracy rates up to 84.1% using linear Support-Vector-Machine classification (lSVM). A permutation test confirmed that all lSVM based prediction rates were significantly better than "guessing". More generally, we present four approaches to analyzing brain function using lSVMs. (1) We show that lSVMs can be used to extract spatio-temporal patterns of brain activation from MEG-data. (2) We show lSVM classification can demonstrate significant correlations between comparatively early and late processes predictive of scene recognition, indicating dependencies between these processes over time. (3) We use lSVM classification to compare the information content of oscillatory and event-related MEG-activations and show they contain a similar amount of and largely overlapping information. (4) A more detailed analysis of single-trial predictiveness of different frequency bands revealed that theta band activity around 5 Hz allowed for highest prediction rates, and these rates are indistinguishable from those obtained with a full dataset. In sum our results clearly demonstrate that lSVMs can reliably predict natural scene recognition from single trial MEG-activation measures and can be a useful tool for analyzing predictive brain function.

Introduction

Human observers extract information from natural scenes at a glance and the memory for them is outstanding (Nickerson, 1965; Standing, 1973). Only a few tens of milliseconds are required to extract essential information from a scene, allowing us to rapidly scan the environment. Brain-networks for analyzing (Thorpe et al., 1996; Grill-Spector et al., 2000; Rieger et al., 2005) and storing (Brewer et al., 1998; Osipova et al., 2006) the content of natural scenes have been described. What remains unknown, however, is if and how reliable brain activation recordings, taken during scene processing, can predict a participant's ability to recognize the scene later on.

Studies using pattern masking to restrict the length of the interval in which information from the scene is available (Gegenfurtner and Rieger, 2000; Grill-Spector et al., 1998; Rieger et al., 2005; Rieger et al., 2008) have revealed that only 60 ms of undistorted processing is sufficient to extract enough information for the reliable later recognition of a scene photograph (Rieger et al., 2005). The information available after these short processing intervals includes scene colors (Gegenfurtner and Rieger, 2000), scene orientation (Rieger et al., 2008), knowledge about objects in the scene (Rieger et al., 2008), and even knowledge about semantic matches between objects and the scene context (Rieger et al., 2008). Recordings of event related potentials (ERP) or event related magnetic fields (EMF) indicate that these scene features are processed relatively early. ERP-signatures of neuronal object processing are found in event related responses between 130 ms and 200 ms after the stimulus onset (Jeffreys, 1996; Thorpe et al., 1996; Allison et al., 1999) and EMFs indicate that within 160 ms of processing basic information about the scene content is extracted (Rieger et al., 2005).

* Corresponding author. Fax: +49 391 67290224.
E-mail address: jochem.rieger@med.ovgu.de (J.W. Rieger).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.06.014

Memory related effects in MEG and EEG have been reported in a latency range longer than basic perceptual processing. Success or failure of memory encoding of visual material is reflected in amplitude modulations starting from 200–300 ms after stimulus onset. These differences may last for several hundred milliseconds (Paller and Wagner, 2002) and have been denoted the "difference in subsequent memory" (DM) effect (Paller et al., 1987). This DM-effect has been extensively studied using event-related potentials, magnetic fields, intracranial recordings, and functional magnetic resonance imaging. The DM-effect with visual material is reflected by modulations within a widespread, dynamic cortical network that includes temporal, prefrontal, and parietal cortices (for reviews see Wagner et al., 1999; Friedman and Johnson, 2000; Paller and Wagner, 2002). Furthermore, recent EEG-studies indicate that success or failure of memory encoding is also reflected in slow wave oscillatory responses (Klimesch et al., 1996, 1997; Sederberg et al., 2003; Osipova et al., 2006; Jensen, 2005). The functional role of these slow-wave oscillations is still under discussion. However, it has been proposed, that these wide spread oscillations reflect the coordinated activation of a distributed brain network during successful memory encoding (Klimesch et al., 1997; Friedman and Johnson, 2000; Ward, 2003; Sederberg et al., 2003; Jensen 2005). Thus, both the event-related responses reflected in time series representation and the frequency representations obtained by decomposing the time series into frequency bands may provide information useful for predicting a participant's recognition success or failure. The exact role of specific frequency bands in recognition remains to be investigated.

Turning to a more methodological issue, studies investigating brain processes involved in recognition typically compare averages of brain activations between conditions (e.g. between recognition failure or success) to detect statistically significant differences in neuronal processing. Although this approach has been successfully applied to investigate neuronal processing, it remains unclear how predictive these statistically significant effects are of trial-by-trial processing. Information about single-trial relevance of brain activation differences is lost because *p*-values of statistically significant outcomes are dependent on the number of measurements included in the analysis, and because the applied significance level is based on an arbitrary agreement. In contrast to statistical significance measures the proportion of correct predictions obtained with single trial classification is a measure of relevance of specific brain activation differences. Single trial classification can provide an answer to the question how well brain networks could discriminate in single trials between different cases based on the informative brain activation patterns retrieved by the classifier. Furthermore, such a classification approach can be used to extract the informative portions of brain activity used by the classifier.

Here we tested whether single trial EMFs allow for accurate predictions of recognition success or failure. Specifically, using linear support vector machines (lSVM) (Vapnik, 1995), a state of the art approach to classification, we aimed to predict from single trial EMFs recorded during the encoding of briefly visible natural scene photographs whether a person will recognize the photograph later on. We used time- and wavelet-derived frequency representations of the data to make single trial predictions because they may highlight complementary functional interpretations of processing differences between recognized and forgotten scenes. Importantly, we tested the reliability of our classification results, extracted brain activations informative about recognition success, and tested how deterministic these processes are over time.

Material and methods

Participants

Seven voluntary participants took part in the experiment after giving their informed consent (4 females, 3 males, mean age 24.6 years). The experiment was in compliance with the Declaration of Helsinki. All participants had normal or corrected-to-normal visual acuity and were paid for their participation.

Stimulus presentation and psychophysical paradigm

Photographs of natural scenes were presented with a Liesegang model ddv810 DLP-projector running at 72 Hz refresh rate. The projector was located outside the MEG shielding chamber and rear projected through a waveguide via a mirror system onto a screen placed at 1.2 m distance from a participants head.

Each experimental trial began with a fixation cross shown for a random duration between 1000 ms and 1400 ms (Fig. 1). Then a photograph of a natural scene was presented for 37 ms and immediately followed by a pattern mask that remained on the screen for a random duration between 1000 ms and 1400 ms. After the mask a red and a green square were presented. Participants were instructed to rate at this time whether they were confident to recognize the scene. They judged they would recognize the scene by lifting the finger on the side of the green rectangle, and judged they would fail by lifting the finger on the side of the red rectangle. The finger movements opened a light barrier, and a signal was recorded in parallel with the MEG. The confidence rating was delayed and jittered with respect to the initial encoding phase to temporally decouple the scene encoding and confidence rating periods. Participants were instructed to rate their confidence of recognizing the scene only after the appearance of the colored squares. After the confidence rating was obtained, four different scenes were presented simultaneously. One of them was the previously presented target scene. Participants indicated the location of the correct target by lifting the finger assigned to its position (four alternative forced choice (4AFC) delayed match-to-sample task). The scene mask stimulus onset asynchrony (SOA) was selected to be short enough to produce a sufficient number of trials in which the participants succeeded or failed to recognize the scene (Rieger et al., 2005).

Correct labeling of trials is extremely important for the successful training of a classifier and for the evaluation of the classification results. Therefore, we included only those trials into the analysis in which the participant's confidence rating was concordant with the later recognition success (class correct: judgment and response correct, class false: judgment and response false). The reasoning behind this selection criterion was that with these trials the likelihood was greatest that estimations of success were based on an evaluation of internal processing. Inconsistent predictions had a greater likelihood of being due to guessing or erroneous finger movements. They were therefore less reliably related to

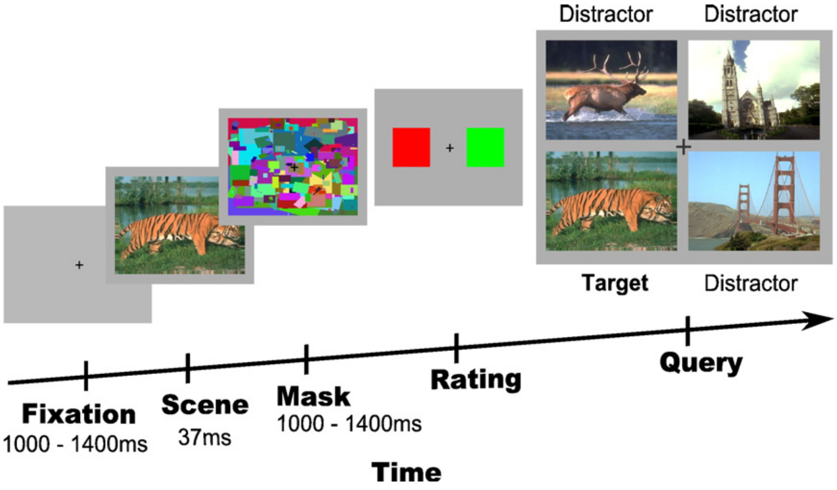

Fig. 1. The figure depicts the psychophysical paradigm. After a variable fixation interval (1000–1400 ms) a photograph of a natural scene was briefly presented (37 ms) and immediately replaced by a mask which remained on the screen for 1000–1400 ms. Next, a red and a green square appeared. Participants were instructed to judge their confidence in recognizing the scene only after the appearance of these squares. They indicated they were confident in subsequently recognizing by lifting the finger on side of the green square and lifted the finger on the side of the red square if they judged they would fail. After this judgment, four scene photographs were presented and the participants used finger movements to indicate the position of the previously presented target. The MEG-recordings used to predict the participant's recognition success on each trial started at the onset of the scene and lasted for 600 ms.

processing differences and less reliable in indicating actual recognition success or failure.

MEG recording and preprocessing

MEG was recorded with a CTF-Omega 151 channel first order gradient system at 625 Hz, digitally low-pass filtered at 40 Hz and then down sampled to 120 Hz. The filtering effectively suppressed the noise from the power line (50 Hz) and from the projector (72 Hz). The higher sampling frequency during recording served to avoid the aliasing of higher harmonics into the low frequency band we analyzed. A photo diode was placed at the upper edge of the screen where a small rectangle switched from black to white simultaneously with the presentation of the target scene. The signal from the photo diode was sampled parallel with the MEG and used to segment the MEG data around the time of the occurrence of the target scene. Epochs with 800 ms duration starting 200 ms prior to scene onset were extracted for analysis. The baseline was calculated from the 200 ms pre-scene interval and the 600 ms interval starting from scene onset was used for classification. Epochs containing artifacts exceeding 3 pT (peak-to-peak) were rejected. The data from one sensor had to be excluded due to frequent malfunction. Each epoch used for classification consisted of 10950 samples (150 sensors × 73 samples).

Classification

We estimated the classification performance generalization by using leave-one-out cross validation (LOOCV). In LOOCV one trial is held out and the classifier is trained on the remaining $n-1$ trials. Then the trained classifier is used to predict the class label of the excluded trial. This procedure is repeated n-times. Finally, the class labels assigned by the classifier are compared to the experimentally obtained class labels to calculate the estimated correct prediction rate. In each LOOCV-iteration test data and training data are strictly separated. This prevents the inclusion of information about the test data into the classifier. Otherwise, the estimated classification performance would be biased towards good classification at the cost of generalization of performance. LOOCV provides an almost unbiased estimate of the generalization error (Lunts and Brailovskiy, 1967; Joachims, 2002) because leaving out one example produces only a small change in the training data set. The disadvantage of LOOCV is the high computational cost entailed by the n trainings of the classifier. Although we tested four different classifiers (SVMs, decision trees, naïve bayesian, and a simple similarity classifier) we report results only from lSVM-classification, because this classifier led to the best and most reliable classification performance.

Guessing level estimation and permutation tests

In a next step classification results were compared to the guessing level to evaluate their reliability. This comparison should reveal whether classification performance obtained with the measured dataset is based on information provided by the data or based on guessing. But how can we obtain a good estimate of the guessing level?

An often used theoretical estimate of the guessing level is the reciprocal of the number of classes in the classification problem. In our study with two classes this would predict the guessing level to be at 50% correct classifications (perfect coin flip). However, other factors might also have an influence on the guessing level. In experiments with probabilistic outcomes the two classes are likely to contain an unequal number of trials. A classifier that only learns the relative frequencies of two class labels is expected to converge towards the theoretical guessing level P_{guess} (see Appendix A for derivation):

$$P_{\text{guess}} = P(c)^2 – P(f)^2 \tag{1}$$

Here $P(c)$ is the probability that the scene was recognized in a trial and $P(f)$ is the probability that the participant failed to recognize the scene (false trial). A third alternative is for the classifier to assign all trials to the class holding more trials if no other information is available. In this last case one would expect the guessing level to correspond to the proportion of trials in the larger group. With equal class sizes all three guessing strategies predict a 50% guessing level. However, with unequal class sizes the last two, class size dependent, strategies would result in guessing levels exceeding 50% correct predictions. In addition, other effects that are either hard to consider analytically, or that may have gone unnoticed could have an influence on the guessing level. Empirical estimates of the guessing level are more likely to capture such effects.

Therefore, we used a permutation procedure to calculate multiple empirical estimates of the guessing level. From these estimates we derived two parameters: The *mean* guessing level and the *confidence interval* of the guessing level. In this permutation procedure class labels were permuted among the training data (i.e. the single trial MEG-measurements) in a dataset. In the next step a full LOOCV was performed on the permuted training set to obtain one guessing level estimate. Permutation and LOOCV were repeated 500 times for each participant's dataset, if not otherwise stated. The resulting 500 estimates of the guessing level were used to estimate the individual mean guessing level and its 95% confidence interval. The prediction rate obtained with the actually measured combinations of class label and single trial MEG-activation should exceed the 95% confidence interval of this estimate. Only then we can assume that the classifier learned from the data and that the results are meaningful (Good, 2005). When all class labels and training data combinations are tested this approach corresponds to Fisher's exact randomization test (Fisher, 1935) which provides an exact 95% confidence interval for the dataset being evaluated. Here, each permutation estimate was based on between 144,000 and 208,500 single trainings of the classifier and corresponding classifications, depending on the total number of trials in a dataset. Importantly, in the permutation approach the guessing level is understood as a random variable varying for different combinations of class labels (recognized/failed) and single trial measurements. Given the fact that empirical data are noisy and contain a limited number of training sets this assumption appears to be more realistic than assuming one fixed guessing level. Furthermore, in contrast to theoretical estimates of guessing levels the permutation estimate is non-parametric.

Here, we report empirical guessing level estimates and compare them to guessing levels expected from theoretical considerations.

Support Vector Machine classification and retrieval of informative brain activation

SVM classifiers are applied in many machine learning problems. Recently, SVMs have been used for single trial classification of fMRI (e.g. Cox and Savoy, 2003; Mourão-Miranda et al., 2005; Haynes et al., 2007), EEG (Hinterberger et al., 2003) and MEG data (Guimaraes et al., 2007). Here we used a publicly available SVM matlab-toolbox (http://ida.first.fraunhofer.de/~anton/software.html) for single trial classification.

SVMs are known to have good classification and generalization performance because the classifier's complexity does not depend on the complexity of the feature space. This independence is especially important when classifying high-dimensional brain imaging data (Cherkassky and Mulier, 1998). In particular, SVMs can implement linear and non-linear classification boundaries by using non-linear kernel functions to transform the data from feature space data into a high dimensional classification space. Here, we only report results from linear classification in the original feature space for two reasons. First, initial tests indicated that the use of higher order polynomial or radial basis function kernels did not increase classification but reduced the generalization performance. Second, our aim was to extract the informative features from the classifier for the analysis of brain function. This is not directly possible with non-linear kernels, although approximative methods have been suggested (Schölkopf et al., 1999). Training a linear classifier provides a normal vector $\vec{w}$ defining the orientation of the separating hyperplane in feature space and an offset b of the hyperplane. Once the classification function is calculated the classification of a trial involves a simple dot product. The classification function is $y_i = \text{sign}(\vec{w} \cdot \vec{x_i} + b)$ where $\vec{x_i}$ is the feature vector of the ith trial to be classified, and y_i is the class assigned to the trial based on the sign of the calculation's result. High entries in $\vec{w}$ weight the features in $\vec{x_i}$ stronger than entries close to zero. Furthermore, high $\vec{w}$ entries indicate the directions in which the margin between classes is wider (the criterion optimized during the training of the SVM) and therefore in the directions of feature space that allow for the best separation. The absolute value of the entries in $\vec{w}$ can be interpreted as a measure for the information a feature provides for classification. If the classification is performed on the time series data each entry in $\vec{w}$ represents the weight for the reading from a specific sensor taken at a specific time. Therefore, the entries in $\vec{w}$ can be visualized as a time series of topographies of informative brain activation similar to regular MEG-time series topographies.

Measures for the evaluation of the classification results

We evaluated the classification performance with three different measures derived from the recognition/classification contingency table (Table 1).

The first measure, proportion of correct predictions, is the percentage of all trials the classifier correctly assigned to the empirically observed recognized/failed classes:

$$P_{\text{correct predictions}} = P(c \cap c') + P(f \cap f') \tag{2}$$

The symbols c and f denote class labels “correct” and “false” assigned empirically (i.e. the measured label). The symbols c' and f' denote the class labels assigned by the classifier (see also Table 1). The proportion of correct predictions is the outcome reported in most studies using classification and has a simple interpretation, but, as previously noted, the guessing level may deviate from 50% depending on the relative group sizes.

Table 1
Recognition/classification contingency table

		Classified as	
		Correct	Failed
Recognition	Correct	$P(c \cap c')$	$P(c \cap f')$
	Failed	$P(f \cap c')$	$P(f \cap f')$

The contingency table describes the possible combinations of a participant's recognition success and the class assigned by the classifier.

Recall and precision are two other measures that evaluate whether the classifier exploits group information contained in the data and are insensitive to group size differences. These measures relate to the rows and the columns in the contingency table (Table 1), respectively. Recall is the proportion of trials that belong to a certain empirical class (observed recognition or failure) and were assigned to this class by the classifier. It measures how well the classifier was able to "recall" the participant's recognition from the MEG-measurements, for example the proportion of the set of trials the participant recognized a scene was actually classified as "scene recognized". The recall reported here is the average of the recalls calculated separately for each class according to the following equation:

$$\mathrm{Recall} = \frac{1}{2}\left(\frac{P(c \cap c')}{P(c \cap c') + P(c \cap f')} + \frac{P(f \cap f')}{P(f \cap f') + P(f \cap c')}\right) \quad (3)$$

Precision takes a complementary perspective. It starts from a set of trials created by the classifier (predicted recognition or failure) and calculates within this set the proportion of trials in which the classification result is in concordance with the participant's recognition performance. Thus precision provides information about how precise a classifier is in its class assignments. Precision was calculated as:

$$\mathrm{Precision} = \frac{1}{2}\left(\frac{P(c \cap c')}{P(c \cap c') + P(f \cap c')} + \frac{P(f \cap f')}{P(f \cap f') + P(c \cap f')}\right) \quad (4)$$

Note that expected average recall and precision guessing level is 50% independent of relative group sizes for all three guessing strategies previously discussed. It can be shown that this guessing level can be expected for all guessing strategies that assign class labels independent of the experimentally obtained labels (see Appendix B).

Results

Psychophysical performance

Table 2 summarizes the proportion of trials obtained with all four possible combinations of confidence judgments (sure/unsure) and recognition success (recognized/failed). On average the seven participants correctly recognized the target scene photograph in 65.3% of all trials. In 59.7% of all trials the participants correctly judged that they would be successful at recognizing the target scene, and in 16.9% of the trials the participants correctly judged that they would fail to recognize it. Only these congruent trials were used in the following analysis. In these trials we can be relatively sure that recognition success or failure is a consequence of differences in neuronal processing and less likely due to other causes such as lucky guesses or careless errors. Please bear in mind, that reliable labels are essential for training the classifier and for evaluating the classification results.

Table 2
Portion of judgment/recognition combinations

Participant	Number of trials	Judgment/recognition			
		Sure/ recognized [%]	Sure/ failed [%]	Unsure/ recognized [%]	Unsure/ failed [%]
P1	423	54.0	1.2	30.5	14.0
P2	463	60.0	10.1	11.4	18.3
P3	485	69.4	5.2	13.1	12.2
P4	411	53.3	8.3	14.1	24.3
P5	533	59.7	12.0	9.6	18.8
P6	480	64.8	9.4	18.8	10.8
P7	424	56.6	10.6	12.0	19.8
Average	459.9	59.7	8.1	15.6	16.9
(SE)	(16.5)	(5.8)	(3.7)	(7.2)	(4.8)

The total number of trials per participant and the portion of trials obtained with all possible combinations of confidence judgments (sure/unsure) and recognition success (recognized/failed). The bottom row lists the averages along with the respective standard errors.

Time series single trial classification

Between 288 and 417 trials were available per participant for classification after rejection of epochs containing artifacts and epochs in which the judged and actual recognition success were incongruent. However, these epochs were not equally distributed among the two classes. On average 77.9% of the epochs were recorded in correct trials (sure/recognized) and the remaining 22.1% of the epochs were from false trials (unsure/failed).

SVM-classification on the full dataset

First we used the time series data of all available trials to predict scene recognition in a LOOCV procedure. Here, linear SVM classification correctly predicted recognition success or failure on average in 78.8% of all single trials. The individual correct prediction rates are depicted in Fig. 2A, and the exact values are listed in Table 3 together with the theoretical guessing level calculated according to Eq. (1).

A correct prediction rate of at least 83% was achieved for the three best participants. For each of the seven participants, the proportion of correct predictions exceeded the 95% confidence interval of the guessing level determined by class label permutation.

Furthermore, the guessing levels predicted with Eq. (1) are nearly identical to the empirical guessing levels (Fig. 2A, average 66.6%) for every participant, but deviate substantially from 50%. The guessing levels expected when the classifier assigns all trials to the larger group exceed the empirical guessing levels on average by 11.3% (standard error 1.2%). This suggests that in our data the classifier learned the relative group sizes when the class labels were permuted.

The average recall was 66.2% and the average precision was 68.1%. The average recall guessing level was 49.8% (average 95% confidence interval for guessing 44.8% to 54.9%) and the average precision guessing level was 49.8% (average 95% confidence interval 44.8% to 55.4%). Together these results clearly indicate that the classifier uses information contained in the MEG-time series to predict the participant's recognition success.

However, at this point it is not clear to what extent the elevated guessing levels (due to the different number of trials in the two classes) may have contributed to the high correct prediction rates. This was investigated in a follow-up analysis.

SVM-classification with equal number of training samples

We removed trials from the larger sample until both groups had an equal number of cases. This was done to investigate to what extent unequal class sizes may have increased not only the guessing level, but also the rate of correct predictions. The high correct prediction rates can be assumed to be based on

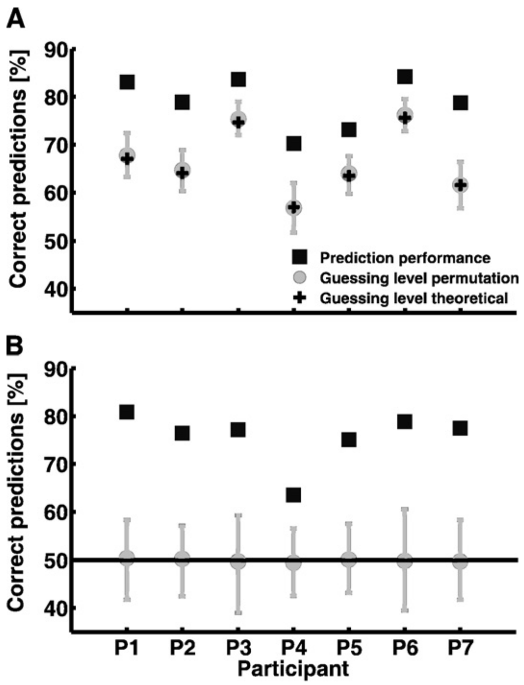

Fig. 2. The figure shows for each participant the recognition prediction accuracies obtained from the time series data using lSVMs (gray squares), together with the results from the permutation estimate of the guessing level (gray disks), the 95% permutation derived confidence intervals for guessing (gray error bars), and a theoretical estimate of the guessing level calculated with Eq. (1) (black crosses in (A) and black horizontal line in (B)). The results in (A) were obtained using the full dataset for training. All correct prediction rates (average 78.8%) exceed the confidence interval for guessing. Theoretical guessing levels fell always close to the permutation derived guessing levels, but deviate substantially from 50% correct predictions. In (B) we discarded trials in the larger training set until equal class sizes were reached. The guessing rate was reduced close to the 50% level, but the correct prediction rate remained nearly unaltered and all prediction rates were significantly better than guessing. This indicates that the good prediction rates obtained with the full dataset were due to information in the training data instead of an elevated guessing level.

information in the MEG-time series rather than on a pedestal contribution from an elevated guessing level if a reduction of the guessing level does not cause a reduction of correct prediction rates (i.e. there is independence of guessing level and correct predictions).

The reduction of the number of training samples led, as expected, to a reduction of the guessing level, which dropped to close to 50% (a reduction of 16.2%). However, the proportion of correct classifications remained nearly unaltered (Fig. 2B, and Table 3). On average the rate of correct predictions was only 3.2% lower (average 75.6% correct classifications) compared to the full datasets. This slight reduction is most likely due to the average 56.3% reduction of the examples available for training the classifier. As with the full dataset the classification performance exceeded the individually determined confidence intervals for guessing. Notably, with equal class sizes recall and precision improved to 75.6%, and 75.7% respectively. Both measures exceed the permutation derived confidence intervals and improved to the same level as the correct prediction rate. These results suggest that the unequal number of training samples contributed only little, if anything, to the recognition prediction we obtained with the full dataset.

In sum these results show that lSVMs can retrieve information from MEG-data for accurate single trial prediction of success of the later recognition of briefly seen natural scenes.

SVM-classification with congruent and incongruent labels

We hypothesized that the selection of trials based on congruent prediction and recognition success contributes to the accurate single trial predictions. We evaluated this hypothesis by determining the prediction performance with the incongruently labeled trials included. In this approach we used the trials with the more reliable congruent labels for the construction of the classifier and tested on all trials in a cross validation. Using this approach the overall prediction accuracy dropped in all datasets compared to using reliable labels during training and testing (average 72.7%, range 65.1% to 78.2%, average 78.8% using only the trials with congruent labels). The reason for this drop was chance level prediction accuracy of recognition success for the portion of trials with incongruent labels (average 49.2%, range 40.2% to 66.4%). Moreover, prediction accuracy further dropped in six out of seven datasets when we repeated the LOOCV ignoring the judgment labels during training and testing (average 70.7%, 63.7% to 80.3%). These results further corroborate our assumption that the labels of incongruent trials are less likely to represent recognition success or failure related to differences in neuronal scene processing. Probably, trials with incongruent labels may provide only very limited useful information for training the classifier and for assessing the accuracy of the classification results.

Table 3
LOOCV results for lSVM trained on the MEG time series

Participant	Full dataset				Equal training class sizes			
	Correct predictions [%]	Guessing level: theoretical (empirical) [%]	Recall [%]	Precision [%]	Correct predictions [%]	Guessing level: theoretical (empirical) [%]	Recall [%]	Precision [%]
P1	83.0	67.0 (67.8)	70.8	74.3	80.8	50.0 (49.6)	80.8	81.0
P2	78.8	64.1 (64.7)	68.2	70.1	76.5	50.0 (49.7)	76.5	76.5
P3	83.6	74.6 (75.4)	63.1	66.2	77.1	50.0 (49.4)	77.1	77.1
P4	70.2	57.0 (56.9)	64.7	65.2	63.5	50.0 (49.4)	63.5	63.5
P5	73.1	63.5 (63.9)	62.5	62.8	75.0	50.0 (49.9)	75.0	75.0
P6	84.1	75.6 (76.2)	63.5	66.2	78.9	50.0 (49.8)	78.9	78.9
P7	78.7	61.6 (61.5)	70.9	72.2	77.4	50.0 (49.4)	77.4	77.5
Average (SE)	78.8 (2.0)	66.2 (2.6) (66.6 (2.7))	66.2/(1.4)	68.1 (1.6)	75.6 (2.1) (49.6 (0.1))	50.0 (0.0)	75.6 (2.1)	75.7 (2.2)

The left four columns list results obtained with lSVMs trained with the full dataset: the percentage of single trial correct prediction, the theoretical (see Eq. (1) in the text) and the permutation test derived (empirical) guessing level for the portion of correct predictions, the recall, and the precision. The same parameters obtained with datasets equalized for the amount of trials in each class are listed in the four columns on the right.

SVM classification on wavelet pyramid coefficients

So far, lSVMs have been trained on the time series of the MEG-sensor readings in this study. However, MEG data are often interpreted in terms of oscillatory brain mechanisms that are represented by certain frequency components in the MEG-data. Moreover, these alternative representations, time series and frequency, may offer different functional interpretations. We were interested in seeing if the classification approach offers different insights when different data representations are used. Therefore, we decomposed the MEG time series into five frequency bands by means of a wavelet pyramid decomposition using the wavelet-toolbox in Matlab R14 (MathWorks Inc, MA, USA). The center frequencies roughly corresponded to those of the five frequency bands which are thought to reflect different functional processes in the brain (Freeman, 1975; Varela et al., 2001; Basar, 2005): The delta band (around 2.5 Hz center frequency), the theta band (around 5 Hz), the alpha band (around 10 Hz), the beta band (around 20 Hz), and the (low) gamma band (around 40 Hz).

Single trial SVM-classifiers trained simultaneously on the wavelet coefficients of all the frequency bands provided results that were nearly identical to those obtained with the time series data. The average rate of correct predictions is virtually identical with the rates obtained with the time series representations (Fig. 3A) and amounted 78.8% (std: 4.5%). The average recall was 66.6% (std: 4.5%) and the average precision was 67.9% (std: 4.8%). Again, all values exceed the individually determined confidence intervals (Table 4).

Table 4
LOOCV results obtained with lSVMs trained on the wavelet coefficients of all frequency bands

Participant	Correct predictions [%]	Guessing level: theoretical (empirical) [%]	Recall [%]	Precision [%]
P1	81.9	67.0 (67.7)	71.4	72.5
P2	79.6	64.1 (64,7)	69.9	71.4
P3	82.6	74.6 (75,5)	63.2	64.7
P4	71.2	57.0 (56.8)	66.0	66.3
P5	72.4	63.5 (63,8)	61.7	61.9
P6	83.3	75.6 (76.1)	61.4	64.0
P7	79.6	61.6 (61,4)	71.9	73.5
Average	78.7 (1.9)	66.2 (2.7)	66.5 (1.7)	67.7 (1.8)

See Table 3 for a description of the parameters in the single columns.

This result shows that the frequency representation and time series representation permitted indistinguishably accurate single trial predictions when all the available information was used. This outcome is compatible with the idea that the two representations provide similar information for classification. We tested this idea further by comparing the classification planes used in both feature spaces. To do this, we transformed the normal vectors $\vec{w}$ of the planes into a common space by applying the inverse wavelet transform to the normal vector obtained in wavelet space. This approach allows for a direct comparison of the information used to predict recognition in wavelet and time series space because after the transform the entries in both $\vec{w}$ correspond to the samples taken at a specific time in a specific sensor. The 3% of the entries that were most informative with these two approaches overlapped by 59.9%. An overlap this high is very unlikely to occur by chance, with a probability of less than 1e−327 (calculated by the odds of binomial coefficients). In sum, the virtually identical prediction rates and extensive overlap of informative dimensions indicate that time series and wavelet derived frequency representations contain a virtually identical amount of information *and* analogous informative features that allow the prediction of single trial recognition success.

The preceding analysis makes use of the full set of wavelet coefficients. However, different frequency bands are thought to reflect different underlying processes, and these processes may contribute differently in the match-to-sample task employed here. We therefore tested specifically whether the delta, theta, alpha, beta and low gamma frequency bands differ with regard to the information they provide about recognition success. The average correct prediction rates obtained separately with each frequency band are shown in Fig. 3B and are listed in Table 5.

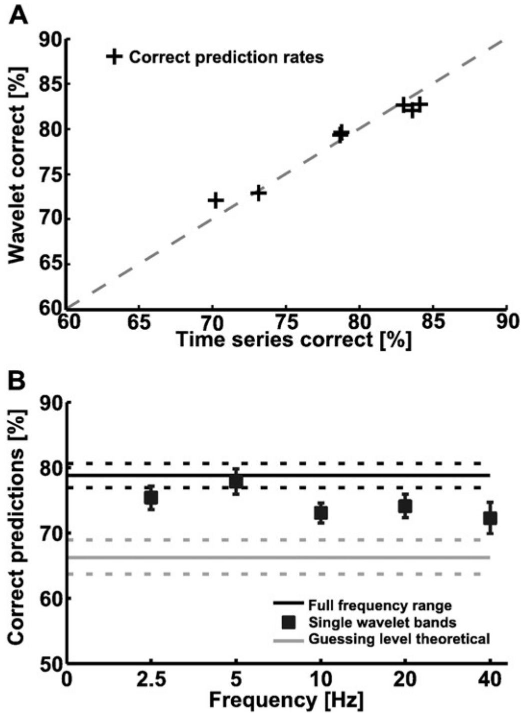

Fig. 3. (A) The figure compares the correct prediction rates obtained with the time series data to the rate obtained with the full set of wavelet coefficients. Therefore, individual correct prediction rates obtained in each feature space are plotted against each other. The results of the two analyses are very similar, and the average rates are virtually identical (see text). The slope of a linear regression does not differ significantly from 1 (r=0.98). Panel (B) depicts the average correct prediction rates (squares) obtained with wavelet coefficients from different frequency bands (delta: 2.5 Hz, theta: 5 Hz, alpha: 10 Hz; beta: 20 Hz and low gamma: 40 Hz). For all frequency bands but the theta band correct prediction rates were lower as those obtained with the full set of wavelet coefficients (black horizontal line, see Table 5 for the statistical tests). This indicates that the theta band is highly informative about recognition success. The gray horizontal line represents the average theoretical guessing level (over participants). Error bars and dashed lines indicate the standard errors of the means over subjects.

Table 5
Average frequency band specific correct prediction rates

	Delta (2.5 Hz)	Theta (5 Hz)	Alpha (10 Hz)	Beta (20 Hz)	Gamma (40 Hz)
Corr. pred. [%]	75.3	77.8	73.0	74.0	72.2
Δ full wc set [%]	−3.4	−0.9	−5.7	−4.7	−6.5
t-value (6 *df*)	3.0	1.3	5.3	6.7	10.4
p-value	0.02	0.24	0.002	<0.001	<0.001

The top row lists the average correct prediction rates obtained using only the wavelet coefficients of the respective frequency band. The second row lists the average difference of band limited prediction rates and prediction rates obtained with the full set of wavelet coefficients (wc). The third and fourth rows list the results of a paired *t*-test comparing band limited and full wc-set prediction rates.

Table 6
Individual theta-band derived prediction rates

Participant	Correct predictions [%]	Guessing level: theoretical (empirical) [%]	Upper 95% conf.int. [%]
P1	80.9	67.0 (67.7)	72.2
P2	77.7	64.1 (67.2)	71.4
P3	83.3	74.6 (77.0)	79.8
P4	69.6	57.0 (57.0)	62.1
P5	75.3	63.5 (66.8)	70.5
P6	83.6	75.6 (76.5)	79.8
P7	74.4	61.6 (62.0)	66.7
Average (SE)	77.8 (1.9)	66.2 (2.6) (66.7(2.7))	71.7 (2.4)

The second column lists the individual correct prediction rates obtained using only the theta band wavelet coefficients for classification. The third column lists the respective theoretical and the mean empirical guessing levels, and the fourth column lists the upper 95% confidence intervals for guessing. The theoretical guessing level was calculated with Eq. (1).

Remarkably, the coefficients from the theta band alone allowed for correct prediction rates that were nearly as high as those obtained using the full set of wavelet coefficients. Using only the theta band coefficients we obtained on average 77.8% correct predictions (Fig. 3B). A set of post hoc *t*-tests comparing individual (n=7) prediction rates obtained with the full set of coefficients to prediction rates obtained with a specific frequency band confirmed a significant reduction of the correct prediction rates for all but the theta band (delta, alpha, beta, and gamma, see Table 5). Moreover, all individual theta band prediction rates were better than guessing in a randomization test (Table 6).

This result has several implications. First, it strongly suggests that a wavelet basis can provide an efficient representation for single trial classification of MEG-data. In our study the theta band requires only 750 wavelet coefficients per trial whilst the full time series includes 10950 samples, and both allow for indistinguishable good classification. This suggests that information about recognition success may concentrate in the theta band. Second, the differences in predictiveness between frequency bands allow for a functional interpretation of the classification results: The fact that the theta band activity permitted the best single trial prediction of recognition success in our delayed-match-to-sample task is concordant with the current view that theta band activity is associated with the encoding of information into working memory (Klimesch et al., 1996, 1997; Paller and Wagner, 2002; Sederberg et al., 2003; Ward, 2003; Osipova et al., 2006). In the next section we investigate another analysis approach which may further contribute to this functional interpretation of the single trial classification results.

The spatio-temporal-structure of information used by the classifiers

In this section we report an analysis of the spatio-temporal MEG-activation patterns that allowed for the good prediction of recognition success using the full time series. The relative predictiveness of the MEG-samples in a trial is represented by the absolute value of the entries in the weight vector $\vec{w}$ of a trained lSVM. Higher absolute values in $\vec{w}$ indicate directions in feature space that allow for better class separation. These feature space directions can be visualized as forming a spatio-temporal pattern because they are linked to readings taken by specific sensors at specific times. In the top row of Fig. 4 we show the 3% most predictive features of the group average weight vector and plotted them as topographic maps. In the bottom row we show the 3% most predictive features of an illustrative participant.

It is evident in Fig. 4 that the most predictive features tend to cluster together. This indicates that the linear SVM captures some of the spatial and temporal correlations produced in MEG-data by temporally and spatially extended brain activations. In an early activation phase, between 100 ms and 200 ms after scene onset, clusters of highly predictive MEG-data concentrate at the occipito-temporal sensors. This suggests that occipito-temporal brain processes involved in visual encoding can already provide information predictive of scene recognition success. Later on, between 200 ms and 600 ms, predictive MEG-activity tends to be distributed at the anterior temporal, parietal and frontal sensors, suggesting activation in a widely distributed brain network that includes temporal, parietal and lateral frontal cortex. Such networks have been associated with the encoding of information into working memory (e.g. Brewer et al., 1998; Takashima et al., 2006; Friedman and Johnson, 2000; Paller and Wagner, 2002). The data from the single participant follow that basic pattern,

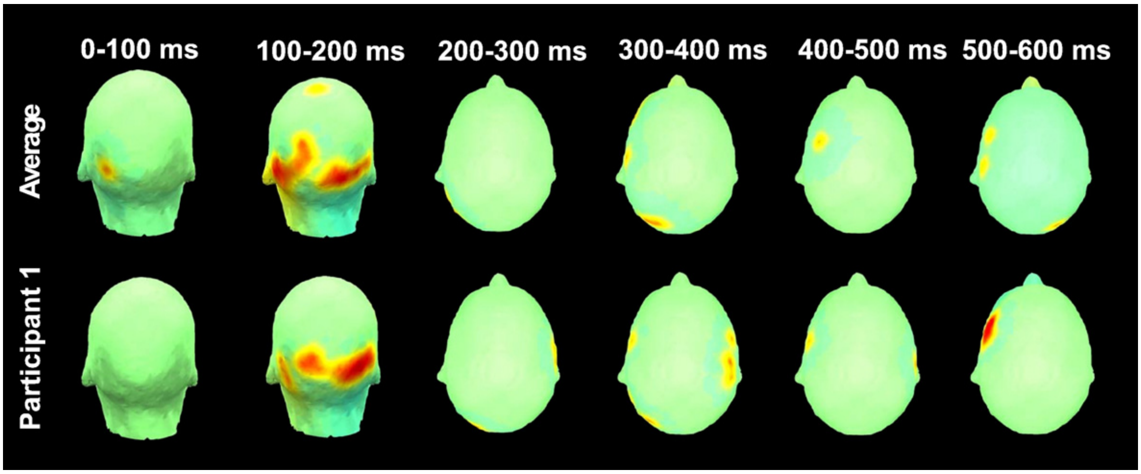

Fig. 4. The figure depicts the group average of the information used by the linear SVM classifiers (top row), and individual data of an illustrative participant (bottom row). Both spatio-temporal patterns are overlaid onto a standard head. The reddish-yellowish blobs the most informative dimensions (highest 3%) for predicting recognition success. During early processing intervals informative data tend to cluster in occipito-temporal sensors, indicating that already relatively early visual processing differences are predictive of recognition success. During a later interval, starting from 200–300 ms, predictive activation was found in more anterior sensors thought to be involved in memory encoding.

although individual deviations are evident. These deviations can be expected because of the strong dependence of the magnetic field topographies on the individual brain anatomy, and because of individual cognitive processing differences.

Together, these results suggest that lSVM allows the retrieval of interpretable spatio-temporal patterns of single trial MEG activity that permit the prediction of recognition success or failure. These patterns indicate that both early, perceptual processing related, and later, memory encoding related brain processes may contribute to recognition success in our paradigm.

Finally we analyzed the coupling of early and late MEG-activation differences.

Analysis of the serial dependence of brain processes

We used the classification approach to estimate the degree of determinism in the sequence of brain processing from the sensory to cognitive and memory formation stages. Therefore, we split each participant's MEG time series in six 100 ms intervals, and performed separate single trial classifications on each interval. This provided six ordered sets of single trial classification results (predicted recognition success or failure in each trial). We then calculated the correlations of the ordered sets of trial-by-trial class labels (−1, 1) between all interval pairs (see Table 7). The rationale behind this approach is that in a strongly determined processing sequence class labels should be more strongly correlated over time than in a weakly deterministic sequence. In other words, we asked whether the single trial recognition predictions derived by classification were related over time.

We found that all pairwise correlations were statistically significant at a Bonferroni corrected level ($p<0.0033$), but the correlations were low. This was especially the case for the correlations between the earliest interval (0 to 100 ms) and the later intervals occurring after 300 ms. These correlations explained less than 0.5% variance. The highest correlations were found between temporally adjacent intervals along the first diagonal in Table 7. However, even the maximal correlation coefficient of 0.28 explains only 7.8% variance. These outcomes suggest that predictions about success or failure are correlated over the processing sequence, and the coupling is somewhat stronger for consecutive compared to non-consecutive intervals. However, given the small amount of variance explained, the coupling over processing intervals with respect to predicted recognition success does not appear to be very tight.

It is conceivable that the relatively low correlations we found are due to insufficient classification performance within the specific intervals rather than weak coupling between brain processes. However, the prediction rates, the average recall, and the average precision are reasonable (Table 8), except for the earliest interval from 0 to 100 ms where all three measures for the classifier's performance are lowest. Moreover, the correlations did not improve when we doubled the interval length (three intervals with a duration of 200 ms each).

Table 7
Correlation of single trial class labels over time intervals

		Later interval				
		100–200 ms	200–300 ms	300–400 ms	400–500 ms	500–600 ms
Earlier interval	0–100 ms	0.17	0.07	0.07	0.06	0.07
	100–200 ms		0.16	0.14	0.15	0.13
	200–300 ms			0.25	0.12	0.16
	300–400 ms				0.23	0.18
	400–500 ms					0.28

The correlations between class labels assigned by lSVMs trained on MEG-activity from different temporal intervals. All correlations are statistically significant at $p<0.0033$ (Bonferroni correction for 15 comparisons).

Table 8
Average LOOCV results for lSVMs trained on 100 ms MEG-intervals ($n=7$)

Interval	Correct predictions (SE) [%]	Recall (SE) [%]	Precision (SE) [%]
0–100 ms	70.0 (2.8)	53.2 (1.3)	54.8 (1.8)
100–200 ms	74.4 (2.1)	60.2 (2.0)	63.1 (2.1)
200–300 ms	74.7 (2.5)	59.9 (1.7)	63.4 (1.7)
300–400 ms	73.8 (2.8)	58.3 (2.3)	61.1 (2.7)
400–500 ms	75.2 (1.9)	62.0 (1.5)	65.1 (1.4)
500–600 ms	77.2 (1.8)	63.5 (1.5)	67.4 (1.4)

Discussion

We have shown that brain activity measured with MEG during the initial processing of a briefly presented natural scene can be used to make reliable trial-by-trial predictions of the subsequent recognition of these scenes. Using a non-parametric randomization test we confirmed that the prediction rates we obtained with lSVM classification are significantly better than guessing. The use of different feature space representations allowed for different but converging functional interpretations of the classification process. The comparison of prediction rates across frequency bands revealed that theta band oscillations, thought to be indicative of memory encoding success, are highly predictive of scene recognition success. The spatio-temporal patterns of brain activation that are informative of recognition success provide additional support for the role of memory encoding processes, but also indicate that early perceptual processing differences may predict success or failure in later recognition tests. Moreover, the analysis of the recognition predictions of successive intervals starting from scene onset indicate significant serial coupling between earlier and later brain processes predictive of scene recognition success. Altogether, our results demonstrate that lSVMs can be used to for reliable prediction of behavior and for the analysis of MEG-measures of brain activation.

Functional interpretation of the informative brain activation used by the classifiers

A central requirement for successful recognition in our delayed-match-to-sample task is the encoding of information about scene content into working memory. Participants are likely to fail to recognize a target scene when either the visual encoding process or the formation of an adequate memory failed. The comparison of predictiveness across frequency bands revealed that theta-band activity was the best predictor of the participant's scene recognition success. Theta band activations allowed for single trial classification performance as good as that observed with the full feature space. Interestingly, the modulation of theta-band activations has been linked to successful memory encoding (Klimesch et al., 1996, 1997; Paller and Wagner, 2002; Düzel et al., 2003; Sederberg et al., 2003; Osipova et al., 2006) and recognition success (Klimesch et al., 2000; Osipova et al., 2006). Theta-band activation is thought to reflect cortical activation induced by cortico-hippocampal feedback loops (e.g. Miller, 1991; Buzsaki, 1996). However, because earlier studies analyzed statistical differences between mean theta activations, it remained unclear how well theta differences could discriminate the success or failure of memory encoding in single trials. Our results provide to our knowledge first evidence that theta band activation can be highly predictive of recognition success in single trials, and thereby emphasize the functional role of theta oscillation modulations in working memory.

The informative brain activation we extracted from time series classification complements and extends this finding. In agreement with previous studies of the DM-effect (for reviews see Paller and Wagner, 2002; Friedman and Johnson, 2000), we found widely distributed clusters of informative frontal, temporal and parietal brain activation. There, long lasting event related MEG-differences that discriminate success and failure in the recognition of pictorial material and words have typically been observed beginning 200 ms to 300 ms after the onset of stimuli in parietal, frontal and temporal sensors (Takashima et al., 2006; Friedman and Johnson, 2000). Like theta band oscillations, these brain activation differences have been attributed to the encoding of information into memory for later recognition and to maintenance in working memory (for a recent review see e.g. D'Esposito, 2007). Despite their wide distribution, these activations must be highly reproducible across trials or we would not have obtained such high and reliable prediction rates.

In addition, SVM-analysis revealed early clusters of predictive activation differences in the interval up to 200 ms after a scene onset in occipito-temporal sensors. The early clusters indicate, that in addition to memory encoding processes, earlier visual processes can be predictive of later recognition success. We think that our use of masked presentations may account for this finding. Pattern masks effectively limit the time available for the extraction of information from visual displays during visual processing (Rolls and Tovee, 1994; Kovacs et al., 1995; Grill-Spector et al., 2000; Rieger et al., 2005, 2008). We have shown in a previous study that employed a similar picture masking paradigm that pattern masks are most effective during the interval between 110 ms and 170 ms after scene onset, when information about scene content is first being acquired (Rieger et al., 2005). This was the same latency range in which we observed predictive MEG differences.

The selection of trials by concordant rating and recognition success may not only improve the reliability of the labels but also raises the question whether brain activation uniquely related to the rating was included into the construction of the classifier. We attempted to temporally decouple scene encoding and rating by instructing participants to rate their confidence only in a rating interval which started 400 ms to 800 ms after the encoding interval used for classification had ended. Despite this experimental precaution, we cannot fully exclude the possibility that participants sometimes judged their success in memory formation prior to the start of the rating interval. Studies in patients (Schnyer et al., 2005) and a recent fMRI-study (Kao et al., 2005) indicate that frontal cortex and more specifically ventromedial prefrontal cortex (VMPC) is involved in judgments of learning. Activation modulations in VMPC due to different confidence ratings would therefore be expected in prefrontal MEG-sensors. However, as shown in Fig. 4 frontal sensors appear to contribute only little information for single trial prediction, indicating that the temporal decoupling was successful. Furthermore, an activation genuinely related to the rating process would only be learned by the classifier if it occurred consistently over trials and time locked to the stimulus. Thus, we think that brain activation related to participant's judgments of recognition success contributed only a little, if anything, to classification performance.

Another question we addressed is how deterministic the processing sequences underlying recognition success or failure are, i.e. to what extend later informative brain activation differences depend on earlier differences. The statistically significant correlations we found between processing intervals suggest that differences in earlier processing intervals determine to some extent the outcomes observed in later processing intervals. However, the small amount of variance explained by these correlations indicates that this coupling may be relatively weak. The weak coupling could either be due to the analysis approach or reflect properties of the brain processes involved in visual processing and memory encoding. We think that the analysis approach is an unlikely explanation because the prediction performance obtained with the short intervals was still good, and because the correlations did not improve when we doubled the length of the analysis intervals. Furthermore, authors relating behavioral decisions to single cell recordings when investigating decision making in monkeys have also concluded there is substantial indeterminacy in the investigated sequence of brain processes (Shadlen et al., 1996; Dorris and Glimcher, 2004; for a review see Glimcher, 2005). Thus, we speculate that the low correlation reflects to some extent indeterminacy in the sequence of brain processes involved in our task. The short scene-mask SOAs we have used in our experiment may have introduced uncertainty that contributed to this indeterminacy. Failures during the early formation of internal representations may have been somewhat independent of failures during the subsequent formation of an enduring trace (Paller and Wagner, 2002). However, further investigations are needed to clarify the causal relationship between brain processes. We suggest that single trial classification may be a helpful analysis tool in these investigations.

Comparison of information about recognition success in frequency and time series representations

Oscillating brain activations are thought to reflect the integration of brain networks at different spatial scales. Different frequency bands have been distinguished and assigned to different cognitive functions (for reviews see Freeman, 1975; Varela et al., 2001; Basar, 2005). On the other hand, event related responses such as time locked deflections in the EEG and MEG time series, are thought to reflect a series of discrete but partly overlapping processing steps. The functional units, which are termed components, are deflections that appear after a relatively fixed latency with a certain scalp topography (reviewed in Rugg and Coles, 1995). Comparisons between these two approaches to the analysis and interpretation of brain function are complex when standard statistical approaches are employed. Studies that have sought to make such comparisons have often had to focus on data from a single sensor, and analyses required many assumptions (e.g. Klimesch et al., 2000; Düzel et al., 2005; Makeig et al., 2002). Single trial classification offers an alternative approach by comparing single trial predictive information among the two types of data representations. The indistinguishably high single trial recognition prediction rates we found indicate that oscillatory and event related brain activations provide a similar amount of information (in the full dataset). However, an equal amount of information does not necessarily imply that the identical information is used by lSVMs in the two data representations. Similar prediction rates could have been achieved with different hyperplanes separating classes within different subspaces of the two data representations. Despite this possibility, the substantial overlap of the informative dimensions we found in the two feature spaces was very unlikely to have occurred by chance. This suggests that the classifier separated classes in similar ways in both the time series and frequency representations of the MEG-data. We therefore conclude that evoked response time series and oscillations provide similar information for single trial classification, at least in our recognition task. This conclusion is further supported by our finding that both the time series and the frequency representations showed memory related brain activation modulations that were predictive of recognition success.

Assessment of classification quality

The large number of data points recorded in each trial during brain imaging creates a potential problem for single trial classification because the number of dimensions in the feature space (data points acquired in each trial) typically exceeds the number of trials available for training the classifier. Therefore, it is essential to assess the reliability of the achieved prediction rates even when these appear high. This has often been done by comparing the empirically obtained correct classification rate to a theoretical guessing level derived from the number of alternatives in the classification problem (e.g. Suppes et al., 1997; Haynes and Rees 2005; Cox and Savoy, 2003; Haxby et al., 2001). Classification rates exceeding this theoretical guessing level are assumed to be significant. Our approach was to use an empirical permutation test for the validation of the classification. Our results strongly suggest that the expected guessing level of the correct prediction rate is not only a function of the number of classes. Empirical guessing levels in our data were much higher than the 50% theoretically expected in a two class problem, indicating that other factors must have an influence. In our study the linear SVM learned the relative class sizes when the class labels were permuted. The good prediction of the empirical guessing level by Eq. (1) is a strong indication for this conclusion. Other classifiers may adopt other strategies that result in different guessing levels (e.g. assigning all trials to the larger class). Unfortunately, unequal class sizes are a frequent problem in cognitive studies in which these sizes are not rigorously bound by the stimulus presentations, such as studies involving spontaneous perceptual switches. Moreover, it is important to note that different class sizes are not the only possible source of bias. For example, temporal correlations in the time series obtained with fMRI may also introduce a bias in the guessing level when single functional volumes are classified. Therefore, we suggest that guessing levels should always be empirically determined, by means of a permutation test. The non-parametric permutation procedure we employed is well suited to estimate the guessing level even when a bias is introduced by unknown sources.

In addition, the permutation method we adopted provides confidence intervals for guessing that can be used for non-parametric significance tests. Only correct classification rates falling outside these confidence bands would be considered significantly different from guessing. Using this procedure we found that single trial prediction rates derived with lSVMs significantly exceeded guessing in every single participant. Importantly, the correct prediction rates achieved with the experimentally obtained class labels were relatively independent on relative group sizes, indicating that unequal class size had only little influence on lSVM classification in our study. One factor contributing to the good classification results we obtained with respect to correct prediction and expected generalization may be the fact that the test error is limited for soft margin classifiers (SVMs) and does not depend on the dimensionality of the feature space (Duda et al., 2001, p.265). In accord with these theoretical considerations, we found that the relatively few theta band coefficients predicted recognition success as well as the full time series despite the 14.6-fold dimensionality reduction of feature space. This is a strong indication that class information in the data rather than the dimensionality of the classification space accounts for the good classification performance we obtained with lSVMs.

Another factor contributing to the good classification results in our study was to select trials in which the participants correctly predicted recognition success or failure. The MEG-data in these trials most likely reflect differences in neuronal scene processing and can be used for analyzing brain function. Conversely, trials with incongruent labels are probably caused by accidental wrong responses or lucky guesses. Thus recognition success or failure assessed by the button press is less likely to be correlated with the neuronal processing recorded during these trials. This assumption is corroborated by our finding that performance dropped when trials with incongruent labels were used for classification and that prediction accuracy for trials with incongruent labels was at chance level. One may argue that selecting trials with congruent labels introduces a bias towards higher accuracy because clear cases are selected. However, reliable labels (labels that correlate with processing differences) are desirable on at least two levels of the classification process to obtain reliable results. First, trials with labels assigned by unclear causes are not useful to train a classifier, and would prevent the interpretation of the classifier and the classification results. This notion is in line with the results of permutation test and with the results obtained with all trials included in the analysis. Second, testing a classifier's accuracy on data with unclear labels may provide less interpretable results and will most likely underestimate a classifier's achievable accuracy.

Proportion of correct predictions and statistical significance

The non-parametric test for the significance of the correct prediction rate we used does not imply that single trial classification and standard statistical testing rely on similar principles. The proportion of correct predictions provided by the classification approach is a relatively simple measure of the relevance of measured brain activity that is hard to derive from results of statistical significance test, e.g. a statistical comparison between average activations measured in different experimental conditions. The reason is that statistical significance depends on the standard error of the mean (SEM) instead of the population variance of the measurements. In theory, the size of the SEM decreases with an increasing number of trials, leading to increasing significance levels for a fixed difference unequal zero. On the other hand, with an increasing number of trials, the estimated population variance converges on the value of the underlying population. It is important to note that the population variance, not the standard error of the mean, is a limiting factor in single trial predictions. Thus, it can be expected that single-trial-classification will converge with an increasing number of examples towards a fixed prediction rate (most likely below 100%). Conversely, with conventional statistical testing small differences that could be of little relevance when a brain must solve a task on a trial-by-trial basis will exceed significance when the number of examples included is sufficiently large. Better predictiveness of task outcomes in classification can be interpreted as signifying higher relevance of a portion of brain activity for solving the task, given that the classifier uses relevant information contained in the data. Viewed in this manner, the proportion of correct predictions has a relatively simple interpretation.

Conclusions

Our results show that it is possible to use linear Support-Vector-Machine Classification to accurately predict a human observer's ability to recognize a natural scene photograph from the first half second of brain activation following the presentation of the scene. Randomization tests provide a relatively simple although computationally intensive way of validating the classification results. Furthermore, we demonstrated four ways to extend the classification approach to analyze the interplay of brain states with behavior using MEG-data: The comparison of the predictiveness of different frequency bands and different feature space representations, the extraction of spatio-temporal patterns of informative brain activation, and an approach to investigate the coupling of brain processes predictive of scene recognition success. Our results are consistent with and extend studies using standard statistical approaches. We conclude that single trial classification is a promising approach for analyzing brain networks predictive of behavior. Moreover, classification provides an easily interpretable measure of the relevance of the informative brain activations: the proportion of correct predictions.

Acknowledgments

We thank Robert Fendrich for helpful comments on the manuscript. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (RI 1511/1-3), and a grant from the Bundesministerium für Bildung und Forschung (FKZ: 1GO0504, Center for Advanced Imaging).

Appendix A. Theoretical guessing level for a classifier learning only the relative class sizes

We assume that the classifier learned only the relative frequencies of the class labels in the training data and assigns class labels with the same relative frequencies during classification. After classification each trial has two class labels: One label was assigned during data-acquisition and the other during classification. The expected guessing level of the classifier is the proportion of trials that received the same class labels in both processes.

We begin the derivation of the guessing level by assuming a first step (the data acquisition) in which a Bernoulli-process assigns class labels c with probability $P(c)$ and class labels f with the probability $P(f) = 1-P(c)$ to single trial MEG-time series. Then a second, independent Bernoulli-process (the classifier) assigns a second class label c' or f' with probabilities $P(c')$ and $P(f')$ to the MEG-time-series. Four combinations are possible:

$$P(c \cap c'); P(f \cap f'); P(c \cap f'); P(f \cap c')$$

and the guessing level is:

$$P_{\text{guess}} = P(c \cap c') + P(f \cap f') \tag{A.1}$$

We assumed the classifier learned the ratio of class labels from the measured data and uses this information to assign class labels with same probabilities. Thus:

$$P(c) = P(c') \tag{A.2}$$

and

$$P(f) = P(f') \tag{A.3}$$

Because both processes are independent

$$P(c \cap c') = P(c)^2 \tag{A.4}$$

and

$$P(f \cap f') = P(f)^2 \tag{A.5}$$

The theoretical guessing level is then:

$$P_{\text{guess}} = P(c)^2 + P(f)^2 \tag{A.6}$$

Appendix B. Guessing level for average recall and precision

Here we derive that the expected average recall and precision guessing level is independent of relative group sizes and independent of the applied guessing strategy. Similar to Appendix A we define guessing the process of assigning class labels independent of the experimentally obtained labels. In other words, the process of assigning a label during data-acquisition is completely independent of the process of assigning a label by a guessing classifier. In this case, the joint probabilities in Eq. (3) can be calculated as a product:

$$\text{Recall} = \frac{1}{2}\left(\frac{P(c) \cdot P(c')}{P(c) \cdot P(c') + P(c) \cdot P(f')} + \frac{P(f) \cdot P(f')}{P(f) \cdot P(f') + P(f) \cdot P(c')}\right) \tag{B.1}$$

Since the probability $P(f')=1-P(c')$ we can substitute $P(f')$. Moreover the probabilities $P(c)$ and $P(f)$ vanish:

$$\text{Recall} = \frac{1}{2}\left(\frac{P(c')}{P(c') + 1-P(c')} + \frac{1-P(c')}{P(c') + 1-P(c')}\right) \tag{B.2}$$

The expression in parenthesis is 1 and therefore

$$\text{Recall} = \frac{1}{2} \tag{B.3}$$

This is the average guessing level for the recall under the assumption that class labels assigned during the experiment and class labels assigned by the classifier are statistically independent. The expected average guessing level Precision $= \frac{1}{2}$ can be derived in a similar way.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2008.06.014.

References

Allison, T., Puce, A., Spencer, D.D., McCarthy, G., 1999. Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. Cereb. Cortex 9, 415–430.

Basar, E., 2005. Memory as the "whole brain work": a large-scale model based on "oscillations in super-synergy". Int. J. Psychophysiol. 58, 199–226.

Brewer, J.B., Zhao, Z., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1998. Making memories: brain activity that predicts how well visual experience will be remembered. Science 281, 1185–1187.

Buzsaki, G., 1996. The hippocampo-neocortical dialogue. Cereb. Cortex 6, 81–92.

Cherkassky, V., Mulier, F., 1998. Learning from Data: Concepts, Theory, and Methods. John Wiley & Sons, New York.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

D'Esposito, M., 2007. From cognitive to neural models of working memory. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 362, 761–772.

Dorris, M.C., Glimcher, P.W., 2004. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron 44, 365–378.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2 ed. John Wiley & Sons, New York.

Düzel, E., Habib, R., Schott, B., Schoenfeld, A., Lobaugh, N., McIntosh, A.R., Scholz, M., Heinze, H.J., 2003. A multivariate, spatiotemporal analysis of electromagnetic time-frequency data of recognition memory. NeuroImage 18, 185–197.

Düzel, E., Neufang, M., Heinze, H.J., 2005. The oscillatory dynamics of recognition memory and its relationship to event-related responses. Cereb. Cortex 15, 1992–2002.

Fisher, R.A., 1935. The Design of Experiments. Hafner, New York.

Freeman, W.J., 1975. Mass Action in the Nervous System: Examination of Neurophysiological Basis of Adoptive Behavior Through the EEG. Academic Press, New York.

Friedman, D., Johnson Jr., R., 2000. Event-related potential (ERP) studies of memory encoding and retrieval: a selective review. Microsc. Res. Tech. 51, 6–28.

Gegenfurtner, K.R., Rieger, J., 2000. Sensory and cognitive contributions of color to the recognition of natural scenes. Curr. Biol. 10, 805–808.

Good, P., 2005. Permutation, Parametric and Bootstrap Tests of Hypothesis, 3 ed. Springer-Verlag, New York.

Glimcher, P.W., 2005. Indeterminacy in brain and behavior. Annu. Rev. Psychol. 56, 25–56.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., Malach, R., 1998. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Hum. Brain Mapp. 6, 316–328.

Grill-Spector, K., Kushnir, T., Hendler, T., Malach, R., 2000. The dynamics of object-selective activation correlate with recognition performance in humans. Nat. Neurosci. 3, 837–843.

Guimaraes, M.P., Wong, D.K., Uy, E.T., Grosenick, L., Suppes, P., 2007. Single-trial classification of MEG recordings. IEEE Trans. Biomed. Eng. 54, 436–443.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17, 323–328.

Hinterberger, T., Kubler, A., Kaiser, J., Neumann, N., Birbaumer, N., 2003. A brain–computer interface (BCI) for the locked-in: comparison of different EEG classifications for the thought translation device. Clin. Neurophysiol. 114, 416–425.

Jeffreys, D.A., 1996. Evoked potential studies of face and object processing. Vis. Cogn. 3, 1–38.

Jensen, O., 2005. Reading the hippocampal code by theta phase-locking. Trends Cogn. Sci. 9, 551–553.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers / Springer.

Kao, Y.C., Davis, E.S., Gabrieli, J.D., 2005. Neural correlates of actual and predicted memory formation. Nat. Neurosci. 8, 1776–1783.

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., 1996. Theta band power in the human scalp EEG and the encoding of new information. Neuroreport 7, 1235–1240.

Klimesch, W., Doppelmayr, M., Pachinger, T., Ripper, B., 1997. Brain oscillations and human memory: EEG correlates in the upper alpha and theta band. Neurosci. Lett. 238, 9–12.

Klimesch, W., Doppelmayr, M., Schwaiger, J., Winkler, T., Gruber, W., 2000. Theta oscillations and the ERP old/new effect: independent phenomena? Clin. Neurophysiol. 111, 781–793.

Kovacs, G., Vogels, R., Orban, G.A., 1995. Cortical correlate of pattern backward masking. Proc. Natl. Acad. Sci. U.S.A. 92, 5587–5591.

Lunts, A., Brailovskiy, V., 1967. Evaluation of attributes obtained in statistical decision rules. Eng. Cybern. 3, 98–109.

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

Miller, R., 1991. Cortico-Hippocampal Interplay and the Representation of Contexts in the Brain. Springer-Verlag, Berlin.

Mourão-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage 28, 980–995.

Nickerson, R.S., 1965. Short-Term memory for complex meaningful visual configurations: a demonstration of capacity. Can. J. Psychol. 19, 155–160.

Osipova, D., Takashima, A., Oostenveld, R., Fernandez, G., Maris, E., Jensen, O., 2006. Theta and gamma oscillations predict encoding and retrieval of declarative memory. J. Neurosci. 26, 7523–7531.

Paller, K.A., Wagner, A.D., 2002. Observing the transformation of experience into memory. Trends Cogn. Sci. 6, 93–102.

Paller, K.A., Kutas, M., Mayes, A.R., 1987. Neural correlates of encoding in an incidental learning paradigm. Electroencephalogr. Clin. Neurophysiol. 67, 360–371.

Rieger, J.W., Braun, C., Bulthoff, H.H., Gegenfurtner, K.R., 2005. The dynamics of visual pattern masking in natural scene processing: a magnetoencephalography study. J. Vis. 5, 275–286.

Rieger, J.W., Köchy, N., Schalk, F., Grüschow, M., Heinze, H.J., 2008. Speed limits: orientation and semantic context interactions constrain natural scene discrimination dynamics. J. Exp. Psychol. Hum. Percept. Perform. 34, 56–76.

Rolls, E.T., Tovee, M.J., 1994. Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proc. Biol. Sci. 257, 9–15.

Rugg, M.D., Coles, M.G.H., 1995. Electrophysiology of Mind: Event-Related Brain Potentials and Cognition, 1 ed. Oxford University Press.

Schnyer, D.M., Nicholls, L., Verfaellie, M., 2005. The role of VMPC in metamemorial judgments of content retrievability. J. Cogn. Neurosci. 17, 832–846.

Schölkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Muller, K.R., Ratsch, G., Smola, A.J., 1999. Input space versus feature space in kernel-based methods. IEEE Trans. Neural Netw. 10, 1000–1017.

Sederberg, P.B., Kahana, M.J., Howard, M.W., Donner, E.J., Madsen, J.R., 2003. Theta and gamma oscillations during encoding predict subsequent recall. J. Neurosci. 23, 10809–10814.

Shadlen, M.N., Britten, K.H., Newsome, W.T., Movshon, J.A., 1996. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J. Neurosci. 16, 1486–1510.

Standing, L., 1973. Learning 10,000 pictures. Q. J. Exp. Psychol. 25, 207–222.

Suppes, P., Lu, Z.L., Han, B., 1997. Brain wave recognition of words. Proc. Natl. Acad. Sci. U.S.A. 94, 14965–14969.

Takashima, A., Jensen, O., Oostenveld, R., Maris, E., van de Coevering, M., Fernandez, G., 2006. Successful declarative memory formation is associated with ongoing activity during encoding in a distributed neocortical network related to working memory: a magnetoencephalography study. Neuroscience 139, 291–297.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J., 2001. The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2, 229–239.

Wagner, A.D., Koutstaal, W., Schacter, D.L., 1999. When encoding yields remembering: insights from event-related neuroimaging. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 354, 1307–1324.

Ward, L.M., 2003. Synchronous neural oscillations and cognitive processes. Trends Cogn. Sci. 7, 553–559.