Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

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

Reading into the mind's eye: Boosting automatic visual recognition with EEG signals

Nicolae Cudlenco [a,*], Nirvana Popescu [b], Marius Leordeanu [a,b,*]

[a] Institute of Mathematics of the Romanian Academy, Bucharest, Romania
[b] University POLITEHNICA of Bucharest, Computer Science Department, Bucharest, Romania

ARTICLE INFO

Article history:
Received 24 July 2019
Revised 5 December 2019
Accepted 19 December 2019
Available online xxx

Communicated by Dr Guo Daqing

Keywords:
EEG
BCI
Object recognition
Computer vision
Deep learning
Neural networks
LSTMs
CNNs
Gabor filters

ABSTRACT

Classifying visual information is an apparently simple and effortless task in our everyday routine, but can we automatically predict what we see from signals emitted by the brain?

While other researchers have already attempted to answer this question, we are the first to show that a commercially available BCI could be effectively used for visual image classification in real-world scenarios – when testing takes place at a completely different time than training data collection. The task is difficult, as it requires relating the noisy and low-level EEG signals to complex and highly semantic visual categories. In this paper, we propose different learning approaches and show that simpler classifiers such as Ridge Regression with Gabor filtering of the input EEG signal could be more effective than the powerful Long Short Term Memory Networks and Convolutional Neural Networks in this case of limited and noisy training data. We analyzed the importance of each electrode for the visual classification task and noticed that the sensors with the highest accuracy were the ones that recorded brain activity from regions known to be correlated more with higher level recognition and cognitive processes and less to lower-level visual signal processing. The result is also in accordance with research in computer vision with deep neural networks, which shows that semantic visual features are learned only at higher levels of neural depth.

While EEG signals are weaker by themselves for the task of visual classification, we demonstrate that they could be powerful when combined with deep visual features extracted from the image, improving performance from 91% to over 97% in a multi-class recognition setting. Our tests show that EEG input brings additional information that is not learned by artificial deep networks on the given image training set. Thus, a commercially available BCI could be effectively used in conjunction with a deep learning based vision system to form together a stronger visual recognition system that is suitable for real-world applications.

1. Introduction

From responding physiologically to concrete and physical stimuli to elaborating opinions and viewing future actions and emotions, the brain is responsible for all these amazing actions, which means that there must be signs indicating their existence. For this reason, the BCI potential has attracted the attention of many researchers for a large set of applications [1]. A typical example would be the use of BCI as a new type of controller [2,3]: to move a wheelchair [4] or for "brain typing"[5].

There are also a few studies that focus on the visual information extracted from EEG, for example the approach for automatic image annotation, using a CNN with an EEGNet architecture [6], or the image reconstruction methods [7,8]. Other recent studies [9,10], successfully use neural networks (CNN and Recurrent NN) for EEG classification tasks and confirm their viability for this kind of problems. The authors of another article [11], used Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Multi-Layer Perceptron Artificial Neural Network (MLP-ANN) and Logistic Regression (LR) in their research work to extract the meaningful EEG signal patterns from a large volume of poor quality data having artifacts noises. For EEG decoding and visualization, deep learning with convolutional neural networks has been used [12]. In this paper, we also investigate a deep learning approach in combination

* Corresponding authors.
E-mail addresses: nicolae.cudlenco@gmail.com (N. Cudlenco), leordeanu@gmail.com (M. Leordeanu).

https://doi.org/10.1016/j.neucom.2019.12.076
0925-2312/

with EEG input, using CNNs and LSTMs, for the task of image classification.

Previous works prove that EEG can be successfully used in several applications and that it is possible to extract meaningful semantic data using BCI. In this paper, we focus on the problem of predicting from EEG data the visual classes seen by human subjects and aim to answer the following questions:

1. Can we accurately predict visual classes from noninvasive EEG signals alone? In literature some articles indicate that invasive systems can be used for similar tasks [13,14]. However, very few showed that the relatively weak EEG signals are relevant for high-level visual classification. A set of recent papers [7,15,16] achieved 83% on multi-class visual recognition from EEG signals alone but in their case the training and testing samples were collected in the same continuous recording session. In this paper we study the more realistic scenario when training and testing data are collected at completely different times. This case is much more difficult, but we show ways in which EEG data can be effectively used for image classification.
2. Is visual recognition based only on features extracted from the brain areas traditionally related to vision or is it the result of a more complex process that also involves other areas of the cortex, responsible for higher level non-vision thought? Consistent with previous research [15,17], our experiments suggest the interesting case that vision might go well beyond simple appearance based processing.
3. Can we improve image classification if we use information extracted from EEG signals in conjunction with standard classic computer vision features? We show that EEG, even when they are weaker than visual features extracted directly from the images, are in fact useful for prediction as complementary signal. By capturing different kinds of information, not learned by deep neural networks directly in the image domain, EEG brings additional discrimination power that significantly boosts the classification accuracy.

There are many studies that attempt to decode EEG data using brain-computer interfaces (BCI) for a multitude of tasks and applications. Only a very small fraction focus on vision [7,8,15,17–19], out of which most pose the problem as a recognition task [17–19]. One closely relates to our work [15], by proposing a deep learning approach in order to predict the class of the image seen by a human subject from the corresponding noninvasive EEG. We will refer to this article as state-of-the-art.

More specifically, the authors address the problem of visual classification using recurrent neural networks and achieve an average accuracy of about 83% on the test set [15]. They collected data for each class, per subject, in a burst of 25 seconds, then used the first 20 seconds for training, the next 2.5 seconds for validation and the last 2.5 seconds for testing. While authors of this article [15] obtained high accuracy, they trained and tested on data taken in the same burst. As a consequence their approach suffers from overfitting as shown in another study [20] and confirmed in our experiments (Section 3.2, Table 5), as their model learns noisy signals that are specific to that particular burst and are less related to the actual semantic image class. We train and test on data taken at different times of the day which is important when making predictions based on EEG signals for real world applications. We take a different approach in data processing and use Gabor filtering in order to remove the high frequency signal that is prone to overfitting when used in combination with powerful deep networks (Section 3.2, Table 5). By our novel processing of the system combined with our electrode signal selection mechanism (Section 3.1, Fig. 4) we are able to achieve a competitive performance in the realistic scenario when the test data is taken at a different time than the train data.

We collected a novel image-EEG dataset by using an affordable, industry-level BCI with 14 electrodes and images from six different classes, including objects and different scenes. We collected the training and test data in distinct sessions, separated by a few hours at least. We wanted to better mimic the real-world conditions and to eliminate all possible interference between the training and test data sets. Instead of choosing from the EEG bands (Epsilon, Delta, Theta, Alpha, Beta, Gamma or Lambda) the ones most relevant for the experiments, we project the entire spectrum on the space of Gabor wavelets, across a relatively large range of frequency bands. As we show in experiments, this approach is efficient and robust to overfitting (Section 3.2, Table 5).

Another related paper augments visual features extracted from images with EEG signals [16]. The authors learn a joint encoding of the visual and EEG information with a Siamese network. However, they used the same dataset as [15] - thus suffering from the same limitation in terms of using training and testing data from the same recording session. Different from their work, we show that EEG data can be effectively used to boost visual recognition even in the case when the training and testing sessions are distinct and relatively far apart in time (Section 2.3, Fig. 9).

Thus, the research problem we are facing is relatively new. The main challenge, as seen in our ablation tests (Section 3.1, Fig. 4), is that EEG signals are weak and generally noisy. They are the aggregate result of firings from billions of neurons, each having specific and often local tasks, over relatively large brain areas. On top of that, the process of capturing qualitative data is made even more difficult by the experimental setup, in which the subject is asked to wear a noise-sensitive and often uncomfortable BCI device for a relatively long amount of time, while remaining focused. Despite the obvious difficulties and challenges posed by the problem, this paper makes several contributions, at the intersection of computer vision and brain computer interfaces:

- We propose an unprecedented approach to classify visual classes from EEG data by capturing signals at different frequencies with Gabor filters and obtain an average classification accuracy of 66.76% over all classes and subjects, and a peak accuracy of 96% on specific classes.
- We investigate the viability of using EEG data alongside state-of-the-art deep neural networks and show a significant boost in recognition from 91% to over 97%.
- We investigate the relevance of each electrode input for classification and experimentally confirm that the most relevant EEG signals come from brain areas that are involved in higher cognitive reasoning, not from areas dedicated to early visual processing (e.g. V1).
- We acquired a novel dataset with over 4 hours of EEG recording from 6 different subjects and 6 different visual classes, including 3 object classes and different 3 outdoor scenes, with distinct training and testing recording sessions, which we will make publicly available.

2. Methods

Our work is motivated by the intuition that when a person is visually understanding a picture, she or he is doing much more than just pattern matching. An image is a glimpse into the human mind. Given enough time to focus, an image will summon all the memories and emotions which, combined, shape the semantic concept behind the pixels. We investigate if we can extract descriptors from EEG, which would allow us to accurately distinguish such different concepts that are triggered by visual input.

Starting from this idea, and to find answers to our initial questions (Question 1, 2 and 3), we designed a three-step architecture, which we present in Fig. 1. First (in EEG recording), we collect EEG

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

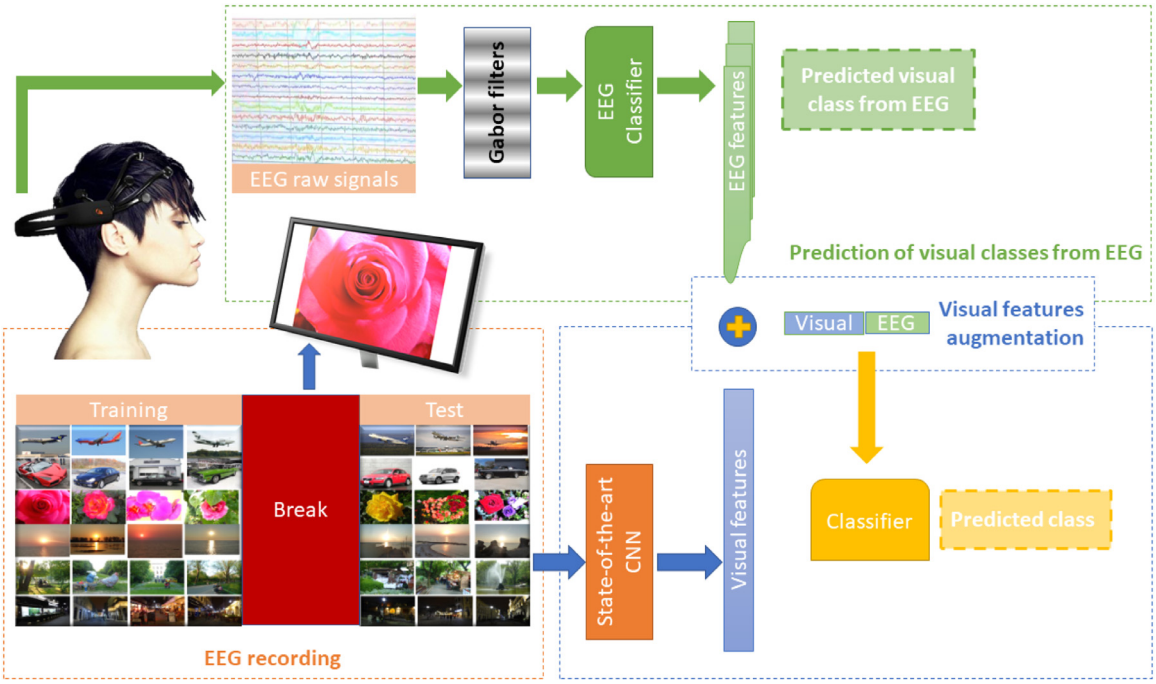

Fig. 1. System architecture. Bottom left: we record EEG signals from people looking at pictures, in distinct training and testing sessions, taken at different times in the day. Top: we apply Gabor filters on the raw EEG data and use it to train an EEG-based classifier. Bottom right: for each picture we extract features from a state-of-the-art deep convolutional net. Then we augment the visual features with the corresponding EEG features learned in the previous step and train a combined Visual-EEG classifier.

data from a group of volunteers. Then (in the prediction of visual classes from EEG task), we apply Gabor filters on the raw data to extract the preliminary descriptors, which we use to train a classifier. In order to better demonstrate the power of EEG for real-world recognition tasks (in the visual features augmentation task), we combine them with visual features extracted with a state-of-the-art CNN (trained from scratch on our images), and feed them to a final classifier for a significant boost in recognition performance.

2.1. Data acquisition setup

To collect the EEG data, a total of six volunteers (4 males and 2 females) participated. All subjects were between 25–35 years old, had similar cultural backgrounds and were all university graduates. We used a commercially available BCI, Emotiv EPOC+[1] with 14 electrodes and 2 reference nodes (CMS and DRL), an internal sampling rate of 2048 downsampled to 128 samples per second (SPS) and a resolution of 14 bits. The electrodes are placed in the International 10–20 System [21] (Fig. 2) and are immobile, they have fixed positions.

Although we are not applying any filters on the data received from the headset (we apply Gabor filtering directly on the raw EEG coming from the BCI), the Emotive EPOC BCI cap applies an internal processing to the data as follows: a strong double notch filter at 50Hz and 60Hz removes interference from the electrical power supply. The filter also affects frequencies down to about 45Hz, so Emotiv specifies 43Hz as the upper usable frequency limit where the spectral response is perfectly flat. The filters extend to about 66Hz, which is higher than the Nyquist cut-off frequency for 128Hz sample frequency.

We chose this relatively inexpensive and therefore easily available equipment, because it has a good quality signal and for that reason it is widely used by the research community. Its fixed electrodes cover relatively uniform the human head and are positioned

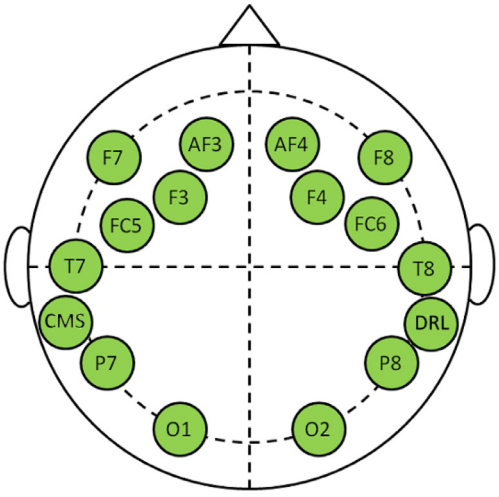

Fig. 2. Emotiv EPOC - electrode placement on scalp. CMS and DRL are the reference electrodes. The locations correspond to the International 10–20 System [21].

according to the International 10–20 System [21] at specific regions of the brain, some of them being placed near the visual cortex and the recognition cortex (O1, O2, P7, P8), others near regions related to other functions. The fact that we do not carefully hand pick the exact positions of the electrodes, makes our results more general and robust to their location. Carefully choosing the electrodes' location might improve performance, but also increase the chance of overfitting. Moreover, the fact that the electrodes cover the entire scalp and are not limited to regions that are known to be dedicated to visual processing (V1) enables us to draw the conclusion that signals from regions that are dedicated to higher level functions could be more relevant for semantic visual recognition than V1.

We obtained an Objects and Scenes EEG dataset, composed of six classes of objects and outdoor scenes: flowers (1), airplanes (2),

[1] Emotiv - brain computer interface technology: http://emotiv.com.

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

cars (3), park (4), seaside (5) and old town (6). The pictures are either selected from ImageNet (cars, airplanes) or were taken by the authors (the rest of classes).

In our experimental setup, the images of a given class are first divided into two groups, one per training or testing session, as follows: 40 images are dedicated for training and 20 for testing. In a given training or testing session the images shown to a subject are grouped per class. Thus, we present the images to subjects in phases, one phase for each of the 6 classes. For example, during the training phase, we start by displaying all training images of the first class on after the other, at a rate of 7 seconds given per image, without pause. Then we move to show the training images of the second class and so on, until all training images of all classes are seen by a given subject. Between classes there is a short pause of around 1 minute. Thus, after showing the images for one class, we wait for a minute and then start showing the images of the next class. Please note that the same protocol is used for both training and testing.

Between the training and testing sessions the subjects take a long break, in the order of hours, during which they take off the BCI device and are encouraged to relax and do other activities. This long break ensures there is no bias between the data collected in the training and testing sessions. It is important to note the significant amount of time passing between the training and the testing acquisition for a given class, with data acquired for other classes in between. In this way we attempted to reduce the noise that could relate the training and testing signals, noise that is not related to the actual image class and could wrongly improve recognition performance (by testing on the training set, in essence).

The chosen order of the classes, which is kept the same for all subjects is the following: 1. Flowers, 2. Airplanes, 3. Cars, 4. Park, 5. Seaside, 6. Old town. Please note that the dataset is quite difficult for the task chosen. The images belonging to a given class vary substantially, in terms of scale, shape, appearance, viewpoint, background scene and number of instances. Thus there is a great variation between the set of images used for training and that used for testing. For a given class we present the images in a preselected, fixed order (chosen randomly at the beginning of the experiments). Thus, all classes and images were presented in the same order for training and testing for all classes and all participants.

During the acquisition process, the images of a given class are shown in sequence, with 7 seconds display time per image. Therefore, for each image class we record 280 seconds for training and 140 seconds for testing. In the pre-processing phase we discard the first 10 seconds to account for the initial setup noise (i.e. the subject finds a comfortable position and focuses on the images); therefore, we use for training the samples from 10 to 280 seconds from the train batch and for test – from 10 to 140 seconds from the test batch. A small subset of the images from the dataset can be viewed in the bottom left side of the Fig. 1.

The signal recorded so far is expected to be noisy, because of the subject's muscles' electrical activity (blinks and other small/micro movements), loud accidental sounds (i.e. outside traffic, ambulance siren), interference from electronic devices or other external factors. To reduce noise we discard the outliers from the resulting EEG. We discarded a sample if on any of the 14 channels it exceeded the channels mean value by a factor of three standard deviations. Three standard deviations were calculated on a subject level, separately for each class. After discarding the outliers from the training data remains an average of 81.77% samples as we show in Table 1. In the end we smooth the data with a small Gaussian filter.

We chose this experiment scenario starting from the idea that an image might trigger, apart from visual processing, other brain activities not necessarily related to vision, like emotions and memories. Our goal is to let the users achieve this target state of mind during training for each class, then, during a long break to let them completely relax and empty their minds so that in the testing phase they would start focusing again to reach the same context of thinking like the one in training. We also consider that this scenario is more suitable for potential applications where we have to train the system once and have test data taken later from a different recording session. For example, in the case of human-machine interaction, the system can first be trained on various concepts, then a BCI could be used to record data continuously, to classify it real-time using our system and perform specific tasks (i.e. send a command to the coffee machine when the user is thinking about coffee).

Table 1
The percentage (%) of remaining training data for each of the six subjects (S1 - S6) after removing the outliers. In average, across subjects, remains a total of 81.77% samples used for training.

S1	S2	S3	S4	S5	S6
82.51	82.25	80.09	83.05	80.79	81.92

2.2. Prediction of visual classes from EEG

We first investigate the idea of predicting image classes from EEG input alone. More specifically, we study whether a low-dimensional but discriminating image class representation is possible, from EEG extracted when the person is looking at the respective picture. Could such a representation be effective for classification?

In existing studies (Ex [22].), the signal is first band passed and only the frequencies of interest (Epsilon, Delta, Theta, Alpha, Beta, Gamma or Lambda) are kept, while all the other data is discarded. Instead, we take an original approach and apply Gabor filters on the raw signal. Each 1D Gabor filter, at a specific scale, could be seen as a band-specific filter. Then, we let the automatic learning process decide the relevance of each filter response.

This approach of using Gabor filters is inspired from computer vision [23] and neuroscience [24]. They are powerful for capturing signal information at different frequencies and time scales - these could be correlated with different cognitive processes and modes of thinking. Gabor filtering also preserves temporal locality which could be important for considering the temporal ordering of human thought. There are also some specific studies in which it is observed that biological receptive fields resemble Gabor filters in the neurological response of cells from the visual cortex [25–27].

Gabor filters are expressed mathematically as a Gaussian modulated by a complex sinusoid. In image processing the two dimensional Gabor filter is used with different frequencies and orientations. In our case we used one-dimensional (1D) Gabor filters with 9 different frequencies - and apply them on each 1D channel. We compute the 1D Gabor descriptors in the following way: the Gabor filter is composed of a real and imaginary part, also known as the even signal, noted here as Es and the odd signal, noted as Os. $Es = \exp(-x^2/(2\sigma^2))\cos((2\pi x)/\lambda)$; $Os = \exp(-x^2/(2\sigma^2))\sin((2\pi x)/\lambda)$, where σ is equal to the value of the wavelength and is the standard deviation of the Gaussian envelope, λ is the wavelength of the sinusoidal factor. The Gabor descriptor is the amplitude $As = \sqrt{Es^2 + Os^2}$. We used 9 wavelength values: $\lambda = 4 \times 1.6^k$, where $k = 0, 1,..., 8$. The resulting preliminary features have a size of $n \times 127$, where n is the number of samples and 127 is the number of dimensions obtained by applying Gabor filters with 9 frequencies on each of the 14 channels (plus 1 for bias, in the case of ridge regression).

In the top part of Fig. 1 we present an overview of the architecture we use for classification. The raw EEG data is first collected and processed with the setup explained in Section 2.1, then the

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

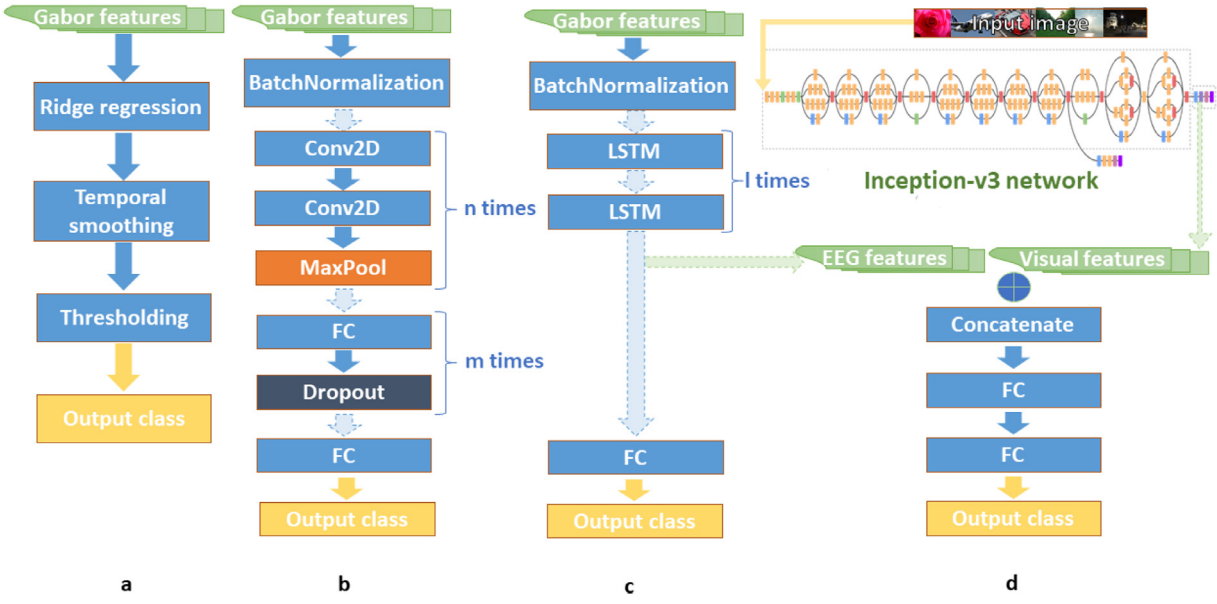

Fig. 3. The classification models proposed. **a)** we use Ridge regression (RR) then apply temporal smoothing on the results and threshold them to identify the final classes; **b)** the CNN architecture - we apply Batch Normalization then have two groups of 2D Convolutional layers with Max Pooling and three groups of Fully Connected with Dropout layers in the end; **c)** the stacked LSTM architecture - we apply Batch Normalization then have one or two LSTM layers with a Fully Connected layer at the end; **d)** the two-step architecture used for visual features augmentation. First we train from scratch the Inception-v3 [28] network on our images and extract for each image its visual features. Then, we train end-to-end the whole network with the visual features concatenated with their according EEG features as input, fed into a Fully Connected layer with ReLu activation. In the final Fully Connected layer we use a softmax activation function. The Inception-v3 architecture image adapted from https://cloud.google.com/tpu/docs/inception-v3-advanced.

preliminary features are extracted with the technique described above and given as input to a classifier. We experimented with multiple classification models:

- *Ridge regression.* (Fig. 3a) A classic and often very efficient machine learning approach using a simple linear classifier. Ridge Regression [29,30] finds the best approximate solution to $X\beta = y$ that minimizes a penalized sum of squares: $RSS + ||\lambda I\beta||_2^2$. By adding a small penalizing factor λ to the diagonal of matrix XX^{-1}, the potential instability of the least squares estimator is fixed. The approximated solution is $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$. To take advantage of the sample's high correlation in time, we apply temporal smoothing on the results of Ridge regression with Gaussian filter. In the end, we compute for each class a threshold which maximizes the accuracy at Equal Error Rate (EER). EER is achieved when the Sensitivity (true positive rate) and Specificity (true negative rate) are close to being equal.
- *CNN* (Fig. 3b) A deep learning approach using convolutional neural networks (CNN). First we create batches of samples recorded each second with a size of (*nFeatures*, 128) (i.e. 128 is the number of SPS; dimension 1, *nFeatures* is 126 padded with 2, for symmetry). The data is normalized with Batch Normalization, then we have two groups (n times) of two Conv2D followed by MaxPooling and in the end three FC with the ReLu activation function followed by Dropout layers (m times). In the end, we have a final FC layer with the number of neurons equal to the number of classes.
- *LSTM* (Fig. 3c) We employ the usage of LSTM layers to better capture the temporal correlation in data. The data is ran first through a BatchNormalization layer. We experimented with one LSTM layer and with two stacked LSTM layers. The input data here is a matrix with all the EEG samples corresponding to an image. The size is 896 $\times$ 128: 896 is the number of secs for an image (7) multiplied with the sampling rate (128) and the number of features ($126 + padded2$).

All deep learning architectures were trained using Stochastic Gradient Descent until an early stopping condition was met. For the one-vs-all approach the dataset becomes unbalanced (small number of positive samples vs high number of negative samples). In this case we train with a penalized loss, inversely proportional to the number of instances in a class. Because of the high class imbalance and low predicting power of the EEG signal, we threshold the output from the last FC layer to maximize the accuracy at EER (Equal Error Rate). For the multi-class case the evaluation is done using the categorical accuracy.

Due to the high variance in data among participants for our experiments we chose to train different classifiers for each subject, similar to other work [15–18]. Training a more generic classifier on the data from all the subjects is a problem we plan to address in our future work.

2.3. Visual features augmentation

To further explore a potential application for brain-computer interfaces (BCI), we test the potential of using the EEG features in conjunction with visual features learned with deep neural nets (CNNs). The general architecture we used can be seen in the bottom-right part of Fig. 1.

We chose a state-of-the-art CNN called Inception-v3 [28] to extract the visual features. The inception networks introduced a new solution to the problem of having large size variation of salient parts in an image. Rather than just stacking convolution operations, a new inception module was introduced, which performs filters with different sizes on the same level. Thus, the networks become wider instead of deeper. For dimensionality reduction the inception module contains convolutions of 1x1. The initial Inception network, GoogLeNet [31], also had two auxiliary classifiers to prevent the problem of vanishing gradients.

Inception-v2 [28] introduced smart factorization (i.e. 5x5 convolutions factorized into two 3x3) and wider filter banks, to reduce the loss of information. In Inception-v3, the authors use 7x7 factorized convolutions, Batch Normalization in the auxiliary classifiers, RMSProp as optimizer and label smoothing. The architecture uses three types of inception modules, as can be seen in Figs. 5, 6 and

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

7 in the original paper [28]. For simplicity, we will refer to them in this article as type A (Fig. 5), B (Fig. 6) and C (Fig. 7).

An onverview of the Inception-v3 architecture can be seen in Fig. 3d. From left to right, in order, there are three conv layers followed by MaxPool, then two more convolutions followed by MaxPool. Then we have three inception model A modules, a grid size reduction bloc, four inception model B modules followed by a grid size reduction block and also connected to an auxiliary classifier. After the last grid size reduction block there are two more inception model C modules, followed by a GlobalAveragePooling layer, Dropout and the final FullyConnected layer with softmax. The details for the grid size reduction block can be found in Figs. 9 and 10 in the original paper [28].

To mimic the conditions when limited data is available for training, we took the Inception-v3 [28] architecture and trained it from scratch on the images used to record EEG training data in our dataset. Thus all EEG-based classifiers and the visual image-based network are trained on data from the exact same set of images. In order to minimize overfitting, the images used for training the Inception-v3 net were also augmented by rotating with a range of 40 degrees, flipping horizontally, zooming with 20%, shifting on the width and height with 20% and applying a shear distortion of 20%. Then, for each image, we extracted the visual features vector from the GlobalAveragePooling layer in the model. Next, we take the EEG data for each image and extract the features from the last LSTM layer in the two-stacked-LSTM architecture (Fig. 3c). Finally, we concatenate the EEG features with the visual features, as shown in Fig. 3d. Then, this joint feature vector is used as input for a simple network with two stacked Fully Connected layers. This model is trained with the joint features for all the images in the training set and tested in the same manner on the images from the test set.

Just as in the previous experiments, we train and test a different classifier, in turn, for each subject. Therefore, a joint feature for an image is composed of the visual features from the Inception net and its corresponding EEG feature taken from one specific subject. Note that the visual features automatically extracted with the Inception module are the same across all subjects - since that is an automated, reproducible process. The EEG signals, however, are taken in turn for each subject and image. We did not concatenate the EEG features from all subjects at once because each brain is expected to process data in its own unique way, specific to each individual, based on experience, genetic factors and many other traits that make each individual unique. Therefore, we do not expect a classifier trained on the EEG signals from one person to perform equally well on a different person.

3. Results

In this Section we perform an in-depth experimental analysis in order to find answers to our initial questions presented in the Introduction (Question 1, 2 and 3). We first start with an ablation study that aims to identify which electrodes (which capture signals from specific areas of the brain) are more relevant for visual classification. We continue with an analysis of the performances of different classifiers in the classification task. Next we present a case study for test and training data collected without the pause and a performance analysis when using EEG in conjunction with visual information.

3.1. Ablation study: Selecting relevant signal

To establish if the EEG signals actually contain information about the visual classes we first looked at the accuracy obtained for each pair of channels using Ridge Regression in a one-vs-all manner. A pair of channels contains two electrodes placed symmetrically on the scalp. We sorted the channel pairs by accuracy (when using only their input for training and classification) in descending order. Then, starting from the pair with the highest accuracy, we created groups[2] by successively adding pairs (of next highest accuracy) until all the channels were used. This is an efficient Greedy approach for studying the effect of different channels and their corresponding brain areas on visual recognition.

It is interesting to note (Fig. 4, Table 2) that even though most of the individual channel pairs have a smaller predictive power by themselves (min. 59.25%, max. 66.21%), the pair with the highest accuracy is very close to the one obtained when using all electrodes together (66.76% with Ridge Regression).

By adding more pairs successively, we achieve a peak of 69.39% (average accuracy on all subjects), after adding the electrodes P7 and P8, placed on the somatosensory association area, responsible for texture, weight and object recognition. The accuracy slightly decreases when adding the channels from the temporal lobe (association area) and frontal lobe (motor functions). In the end, the channels O1 and O2 contribute with some information from the visual area and increase the accuracy.

A surprising result here is the actual order of the channels' relevance. We know from neuroscience research that for each cognitive function or function of the human body there is a corresponding brain activity, which in general can be localized to a specific brain area [32]. While in general there is no single limited area dedicated to a single function, it is also accepted that specific regions are correlated with specific functions, such as higher level thinking or visual processing [32] - and the electrodes, placed on different places on the scalp are expected to capture such correlations. We correlated the physical location of the electrodes on the scalp (Fig. 2) with the brain areas from which they would collect most of the information (Fig. 5). In Table 2, we present the pairs of electrodes sorted descending by their accuracy with their corresponding brain areas (and their indexes from Fig. 5) and the functions of those brain areas.

By analyzing the correlation between channels and brain functions, we can see that the pair with the highest accuracy is the one placed on the area related to the face. The exact placement of the electrode corresponds to the lower part of the cortical motor homunculus, related to the facial muscles [33]. This could be due to specific eye movement patterns (saccades) related to specific classes. It could also be related to the firing of mirror neurons, when the subjects focus on the images and project themselves into the specific story related to the image class, exhibiting motor intentions and facial expressions [34]. More importantly, the signals from area 4 (Brocca's area) are related to speech - which could suggest an internal verbalization of the classes name. Note that each class, either being scene related (park, old town, seaside) or object related (flower, car, airplane) could trigger specific intentions, determining different facial expressions, eye movements, actions or spoken language. Note that the poorer performance of Pair 6 (also related to the motor areas 3 and 12) could be due to the lacking of information from area 4 (speech-related Brocca's area).

The second and third place are the channels located on the higher mental functions area. This suggests that visual perception is a complex process which involves, apart from processing the visual signal, higher brain functions. Only on the fourth place we see the electrodes related to visual object recognition and pattern detection. However, as seen in Fig. 4, they contribute with important information for peak accuracy. It is interesting to see experimentally the connection between semantic visual recognition and different cognitive functions, not all dedicated to direct visual processing. Also note that the ordering of the channel pairs impor-

[2] Group1:Pair1;2 Group2:Pair1;2;3 Group3:Pair1;2;3;4 Group4: Pair1;2;3;4;5 Group5: Pair1;2;3;4;5;6.

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

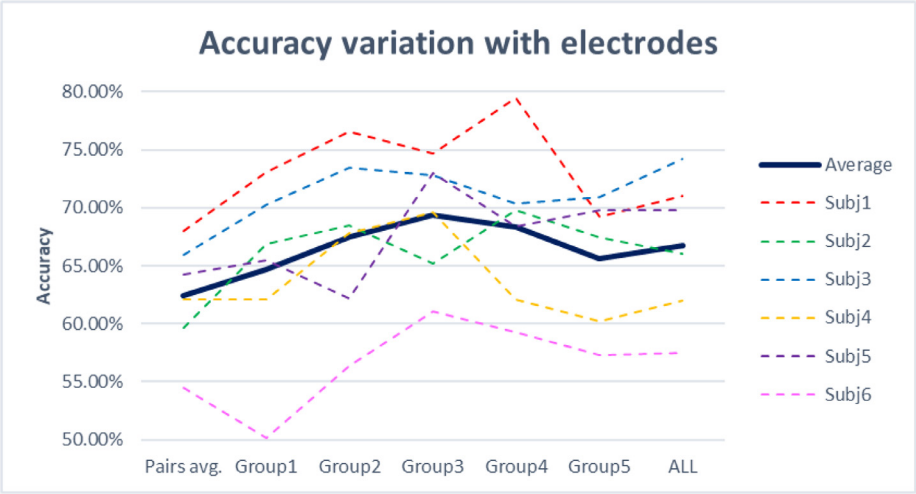

Fig. 4. Accuracy variation with different subsets of electrodes used for ridge regression training and classification. Pairs avg - the average accuracy when using only one pair of channels as input EEG signal; Group1-5^2 - classification using groups of pairs, the pairs in a group are added successively, in decreasing order of their individual accuracy. The electrodes in a pair are presented in Table 2. The accuracy increases when the electrodes placed on the visual cortex are added (Group2 to Group3; Group5 to ALL). The group that generalizes best (Group 3) includes electrodes from areas performing high-level reasoning and visual recognition.

Table 2
Pairs of channels, their accuracy when used alone for training and their corresponding brain areas. The brain area indexes are the ones from Fig. 5.

Channels	Acc (%)	Brain areas and functions
Pair1: FC5;FC6	66.21	*Frontal-central lobe (3,4,12)* 3. Motor function area • Initiation of voluntary muscles 4. Broca's area • Muscles of speech 12. Motor function area: • Face, tongue, larynges muscles • Eye movement and orientation
Pair2: F7;F8	64.14	*Frontal lobe (13)*
Pair3: AF3;AF4	63.02	*Prefrontal lobe (13)* 13. Higher mental functions • Concentration; Creativity • Planning; Judgment • Emotional expression • Inhibition
Pair4: P7;P8	62.94	*Parietal lobe (9,10,11)* 9. Sensory area • Sensations from muscle and skin 10. Somatosensory association area • Evaluation of weight, texture, etc. for object recognition 11. Wernicke's area • Written and spoken language comprehension
Pair5: T7;T8	61.69	*Temporal lobe (2)* 2. Association area • Short-term memory • Equilibrium • Emotion
Pair6: F3;F4	59.66	*Frontal lobe (3,12)* 3. Motor function area Initiation of voluntary muscles 12. Motor function area: • Face, tongue, larynges muscles • Eye movement and orientation
Pair7: O1;O2	59.25	*Occipital area(1)* 1. Visual area • Sight • Image perception • Image recognition

tance is stable across all subjects, which further validates the above observations.

It is important to note here that while we do not directly measure specific brain functions and areas, it is well known that different regions tend to correlate with different functions [32] as shown in our Fig. 5. Therefore, we draw our conclusions based on the noninvasive signal readings from electrodes that are placed in proximity of specific brain regions. We observe not only that signals from higher level cognitive processes are more important for visual recognition than signals from V1, but also that such patterns of evidence is similar across different subjects. This facts suggest that this is a more general observation about where semantic visual recognition might take place in the brain. The idea that visual recognition could involve deeper and more semantic processes is both beautiful, somewhat surprising, but it also makes intuitive sense: in order to understand what an object is, one should also consider the role played by that object in the larger context of a more complex scene in which several actors interplay. Then, putting things in the right context should definitely involve higher-level thinking.

Concluding remarks regarding our empirical observations. The electrodes' locations follow the International 10–20 standard [21]. They cover uniformly the head and are in the vicinity of certain brain regions whose functioning is known to be correlated with specific cognitive functions [32].

Thus, by observing how the signal from different groups of electrodes influences recognition performance we could indirectly infer, empirically, the involvement of different brain functions to semantic visual recognition of objects and scenes in images (Section 3.1).

By doing so, we observed (Fig. 4) a surprising fact: higher level cognitive functions seem to be more relevant for semantic visual recognition (at the category level) than the initial visual processing that is known to take place in V1 area [32].

This finding makes perfect sense: higher level semantic visual understanding at the level of object and scene classes should also require the complex understanding of context, in which object classes are also understood based on the roles they play in the overall visual story. Putting objects in the right context is a higher level cognitive ability, which is expected to require a more global understanding of the scene, achieved by the collaboration of var-

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

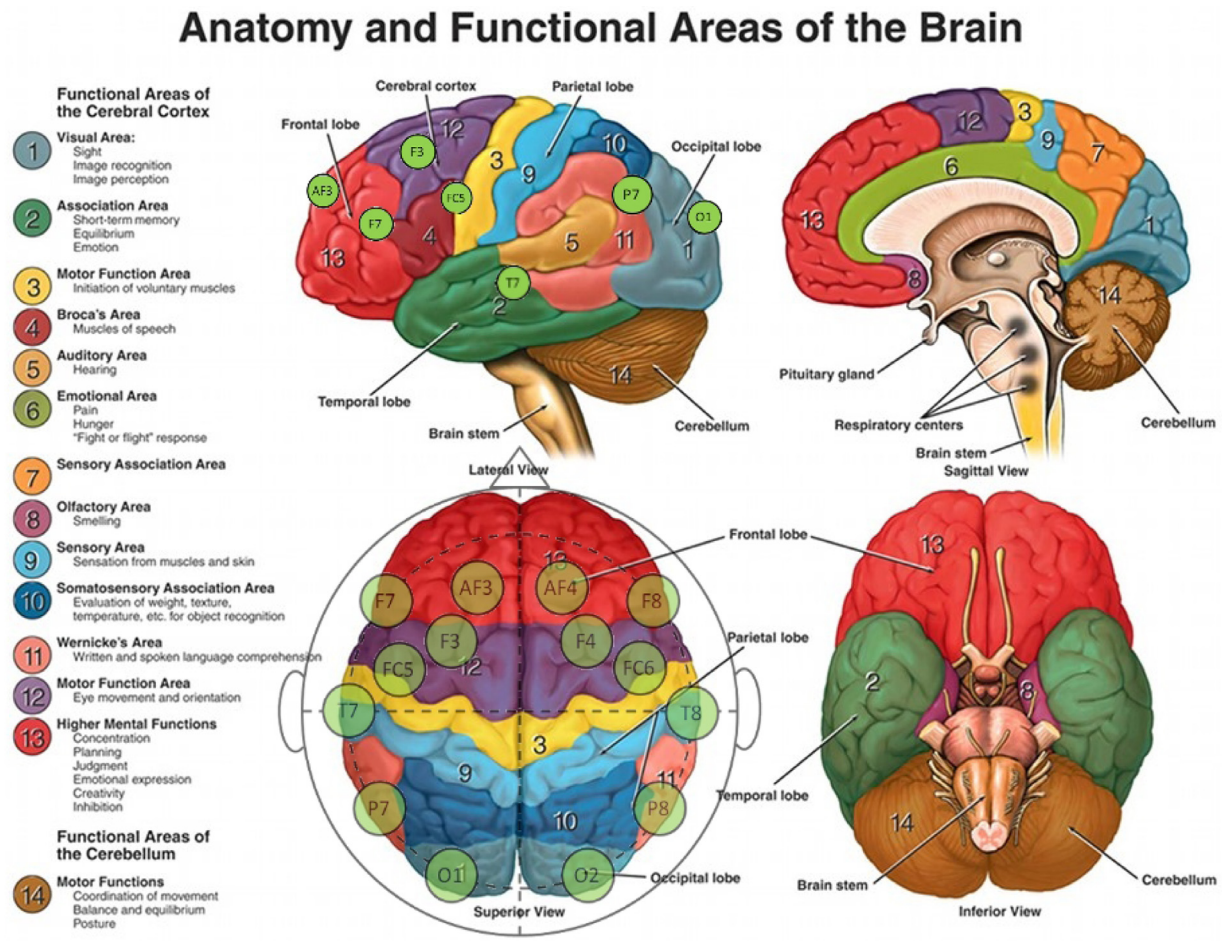

Fig. 5. Electrodes' locations relative to brain areas. Adapted from Neuroanatomy - A Primer, by K. Sukel, 2011, http://www.dana.org/News/Details.aspx?id=43515.

Table 3
Average accuracy (%) per model at EER. The best results are obtained with Ridge Regression (RR). The LSTM architectures appear to be overfitting. All models are trained on all channels, except RR-G3 which is trained on Group3 and is added here for comparison.

	RR	RR-G3	CNN	LSTM-1	LSTM-2
Avg acc.	**66.76**	**69.39**	57.08	58.52	60.97

Table 4
Average accuracy (%) at EER per subject. Ridge Regression results using all channels and only the channels from Group3 (G3). The subjects are sorted in descending order. With **Bold**, *Italic* and ***Bold italic***, in that order, we have the top 3 values. Note the relatively stable ordering among subjects.

	Subj3	Subj1	Subj5	Subj2	Subj4	Subj6
All	**74.17**	*70.79*	***69.82***	66.06	62.01	57.52
G3	***72.82***	**74.71**	*72.99*	65.19	69.61	61.01

ious higher level cognitive processes. They might not be directly involved in the processing of visual input at the lower levels. We could expect however that such higher level cognitive modules take, internally in the brain, inputs from V1 - a fact that cannot be observed by our noninvasive and relatively superficial EEG cues. The validity of our experimental observations is further strengthened by the fact that a similar pattern of electrodes' relevance has been observed across all subjects in our tests (Fig. 4).

3.2. Classification task: Comparisons of models and subjects

We train a different classifier for each subject. The success of this task should mainly depend on two factors: 1) the ability of the classification models to capture and use the relevant information from the EEG signals (when that exists) and 2) the subjects' ability to focus and emit good quality EEG signals. Our experiments show that there is a visible and relatively stable dependency of accuracy on these factors. We experimented with different classification models, as explained in Section 2.2, with a one-vs-all approach. We present in Table 3 the accuracy obtained for each model.

We can see that Ridge Regression outperforms the other models with 66.76% accuracy when trained using all the channels. The stacked LSTM with 2 layers performs better than the CNN, being more suited for time series classification. However, since the dataset is small and the signal noisy, the richer deep models are prone to overfitting, which explains the superior performance of ridge regression. For the rest of the experiments (except Visual features augmentation), we used Ridge Regression with Gabor filtering, since it gives the best results.

In Table 4 we ranked the subjects by their accuracy for the case when we use all electrodes for training and when using only electrodes from Group 3. There is a clear difference in accuracy between subjects. This indicates that the results depend not only on the classification model but also on the subject who generated the EEG. Therefore, we expect that a more trained subject, with a greater ability to focus would generate a richer and better EEG signal and would help the machine learning models learn and predict

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

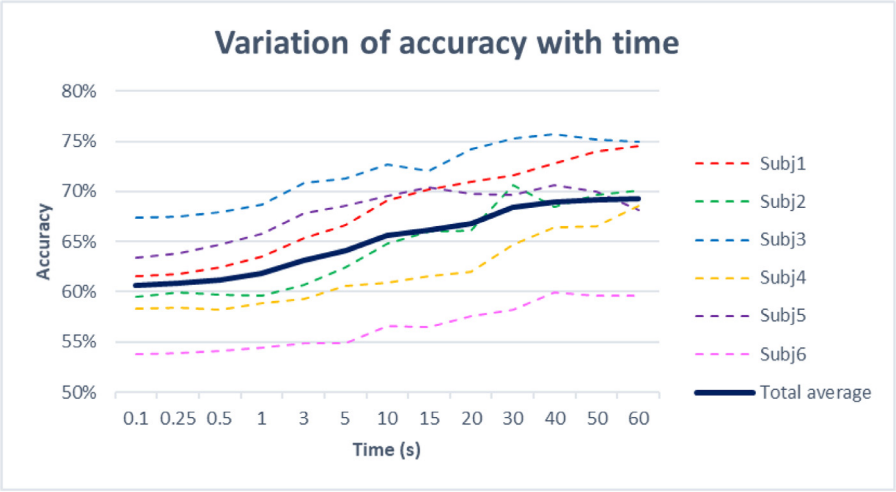

Fig. 6. Accuracy (%) in time. The accuracy increases when considering samples over large time spans.

Table 5
Comparison with the work of [15] - multiclass trained and tested on our dataset. CNN, LSTM-1 and LSTM-2 are our architectures. The accuracy of [15] drops significantly training on our data.

Model	Avg accuracy (%)	
	Raw	Gabors
[15]	18.81	22
RR	20.86	**29.2**
CNN	21.11	26.5
LSTM-1	21.63	25.67
LSTM-2	19.81	22.83

more accurately. Note that the top three when using all channels are the same with the top three in the Group 3 case (Table 4). The result strengthens the idea that Group 3 contains electrodes that are generally relevant for this task, across different people, and can be robustly used for improved classification.

When training the Ridge Regression model we make a prediction for each time sample. Since the signal is actually a time series, we average the outputs over temporal windows of different sizes, centered at the current point. We vary this averaging temporal window (by increasing the corresponding samples taken into account) from 0 to 60 seconds. As expected (Fig. 6) we observe that by taking more samples into consideration, over larger time spans, the accuracy increases consistently.

When analyzing the results for each subject individually (Fig. 7), we see that the class flowers has the highest accuracy for five out of six subjects, with a maximum at 96.35% for Subject 1. For Subject 3 the class with the highest accuracy was cars. The class flowers was the first class in the sequence for which we recorded data. The fact that it appears predominantly on the first place suggests that the subject's fatigue also plays a factor when collecting EEG data. Even more so, the class old town, which was always last in the data collection process, appears to have, for three subjects, the lowest accuracy - which further suggests the subject's fatigue and diminished ability to focus. Another observation is that the subjects claimed to have a special affinity to the class with the highest accuracy. Subject 3 had an intense memory related to the class cars and four of the other subjects (two males and two females) declared a particular affinity for flowers, being related to specific powerful events.

To further understand the difference between our work and the work in the state-of-the-art article [15], we trained and tested the model in state-of-the-art on our Objects and Scenes dataset, preprocessed and Gabor filtered, in our multiclass setting (predict the correct class out of six possible ones) and achieved the accuracy of 22%. In Table 5 we also included the results of our architectures in the same multiclasses setup. Ridge Regression remains the best model, with an accuracy of 29.2%. The random accuracy here is $1/6 = 16.66\%$. Note that on our data, Gabor filtering helps every time, which again suggests that our data is less vulnerable to noisy components that might depend on specific recording sessions. Comparing the results using the model from state-of-the-art and the data obtained with the experiment setup in state-of-the-art [15] (83% accuracy) with the results using the same model but with data from our Objects and Scenes dataset (22%), we notice a drastic drop in accuracy (a drop of 61% when applying Gabors and around 64% when using the raw signal). The results indicate that our experimental scenario, when we essentially allow the "brain-cache" to be erased by introducing a long break between training and testing, is more realistic and therefore more challenging.

3.3. Case study: Collecting training and test data separated vs continuous

In this case study we explore the difference between collecting test and training data in a continuous session versus collecting in different sessions, separated by a long break.

We repeated the process of data collection with three of the subjects (1 male and 2 females), but this time in a different setup (the continuous setup). In this continuous setup, as opposed to the original, separated setup (in which there is a significant amount of time passing between the training and test sessions), we collected the training and test data continuously, without the long pause between the training and testing sessions. We employed the same experimental method as in Section 2.2, using the Ridge Regression model.

As expected, for the continuous data we obtained greater accuracy. We can see in Fig. 8 the results for each subject. The average accuracy is lower in the separated setup and is much higher in the continuous setup. We even have an accuracy of 99.99% on the continuous setup for Subject 6 and the class City. Many factors could stand as grounds for these results. One of them is the fact that the signal has less noise caused by the variation of the electrodes position on the scalp, because in this case the BCI device was not taken off between training and testing. Another factor is the fact that in this scenario the subjects do not have to disrupt their state of mind when proceeding to collect data for the test phase.

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

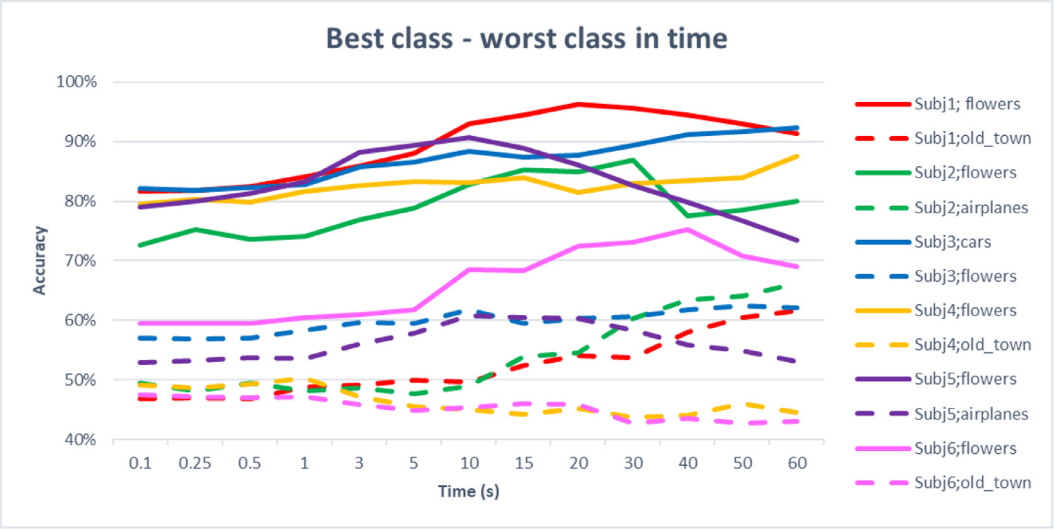

Fig. 7. Best and worst class for each subject in time. Flowers is the predominant best class.

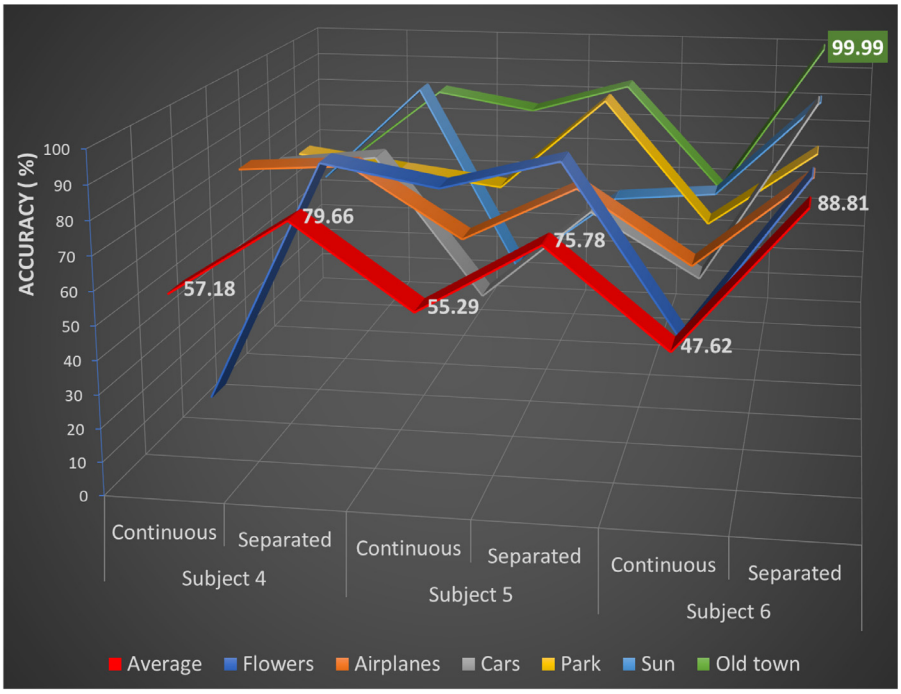

Fig. 8. Continuous vs separated data. The red line represents the average accuracy (%). The accuracy is better when test data is collected immediately after training, with an outlier of 99.99% for Subject 6 on Old town.

When the subjects focus on a image, their brain infers concepts correlated internally to an original set of life experiences, it processes the input data and generates trains of thoughts, stories. Apart from that it may process in parallel other thoughts and activities, unrelated to the concept of interest. We did our best to isolate the signal most relevant for the respective class, to increase the signal to noise ratio by applying the methods discussed in the data acquisition Section 2.1. In this scenario, however, if we do not disrupt the collection session then we could end up learning descriptors specific to this continuous time-span, making the problem much easier but unrealistic.

The conclusion here is that a crucial factor in achieving accurate results with our method is the subject's level of training to reach exactly the same state of mind for a certain class. Another fact to think about is that when there is no time break between training and testing session, then the results could be biased since the subject's brain activity is highly correlated in time. It is very likely that when there is no break between training and testing sessision, we are in fact training on almost the same data as the one used for testing, so we should expect a high level of overfitting.

3.4. Final setup: Boosting visual recognition with EEG

Our final setup is to explore the possibility of enriching the image-based features learned by a CNN with EEG data (Section 2.3). To establish a baseline, we first tested using image-based CNN features only (Fig. 9). Then, we used the augmented image+EEG features and tested for each subject. The EEG signals

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

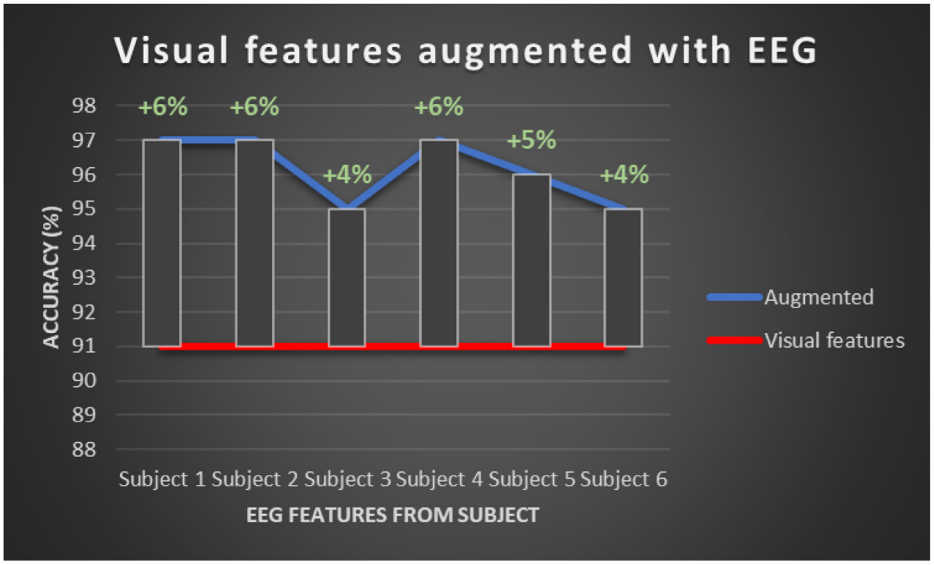

Fig. 9. Visual features augmented with EEG. With red is represented the baseline accuracy, when using only visual information. The other line represent the accuracy for when using visual features augmented with the EEG features from each subject. The average improvement is of 5.16%.

boost performance every time, with a significant overall improvement of 5.16% in the 91+% regime, on the multiclass recognition scenario on our Objects and Scenes dataset.

These results strongly suggest that noninvasive BCI could significantly improve automatic visual recognition, when used in combination with current computer vision and deep learning models.

4. Discussion and concluding remarks

In this paper we tackled the problem of predicting visual classes using only the EEG signal recorded from humans while they're looking at images. Despite the challenges regarding the noisy and hard to collect EEG signal, we presented the following contributions:

1. We proposed and evaluated different automatic models for predicting what humans see by using their EEG signals by themselves (Table 3) or in combination with deep image-based features (Fig. 9). We showed that in the case of noisy EEG data and limited training data, a simpler Ridge regression model combined with Gabor filtering could be at least as effective as the powerful LSTM or CNN models (Table 3). We also conducted an ablation study to determine the most relevant electrodes for classification (Fig. 4).
2. We proposed the use of Gabor filters as a EEG processing technique, which makes possible the removal of noisy signals and improving generalization power (Table 5).
3. We study the involvement of higher level cognitive areas in the visual recognition process and show that visual recognition is the result of the processing in several cortical areas, more relevant being the ones that are involved in higher level cognitive processes. Interestingly enough, the area that is involved in early visual processing (V1) [32] seems to be less relevant - a fact that confirms results from computer vision and deep learning, in which the high level semantic features are learned at the deepest levels of processing (Table 2, Fig. 5).
4. We showed that prediction from EEG data alone is possible but has moderate accuracy, and depends not only on the classifiers but also on the human subjects themselves (Table 4). Our experiments also show that performance strongly depends on the subjects' ability to focus. When they can better project into the "memory spaces" related to specific concepts the recognition accuracy from EEG data alone can go above 95% (Fig. 7).
5. We show that while EEG signals are weaker than visual features extracted directly from images, they contain complementary information which can be used to significantly boost the performance of vision only approaches with deep neural nets. This opens the door for many applications of EEG in the field of visual recognition for human computer interaction (Fig. 9).
6. We propose a **more realistic** experimental setup for data collection, when training and testing data is collected during entirely different recording sessions, at different times of the day and we show by comparing with previous work how this scenario, though more challenging, eliminates the risks of having in the EEG signal noise specific to the recording session (Table 5).

Our results guide our future research. Next, we aim to capture better quality EEG signals by studying the effect of subjects' training and improvement of their ability to focus. Second, we will explore different territories for using EEG to boost automatic visual recognition and vision to language translation. Ultimately, we hope that our work will open doors towards better comprehending human "perception" in relation to "meaning" and the overall understanding of the scene, in a multidisciplinary research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Nicolae Cudlenco: Visualization, Writing - review & editing, Writing - original draft, Data curation, Investigation, Formal analysis, Validation, Software, Methodology, Conceptualization. **Nirvana Popescu:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing - review & editing, Supervision. **Marius Leordeanu:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076

Acknowledgements

This work has been supported in part through UEFISCDI, under EEA-Grants project EEA-RO-2018-0496 and project PN-III-P1-1.1-TE-2016-2182.

References

[1] M. Ahn, M. Lee, J. Choi, S.C. Jun, A review of brain-computer interface games and an opinion survey from researchers, developers and users, Sensors 14 (8) (2014) 14601–14633.
[2] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain–computer interfaces for communication and control, Clinical neurophysiology 113 (6) (2002) 767–791.
[3] R. Krepki, B. Blankertz, G. Curio, K.-R. Müller, The berlin brain-computer interface (bbci)–towards a new communication channel for online control in gaming applications, Multimed Tools Appl 33 (1) (2007) 73–90.
[4] S. Amiri, R. Fazel-Rezai, V. Asadpour, A review of hybrid brain-computer interface systems, Advances in Human-Computer Interaction 2013 (2013) 1.
[5] C. da Silva Souto, H. Lüddemann, S. Lipski, M. Dietz, B. Kollmeier, Influence of attention on speech-rhythm evoked potentials: first steps towards an auditory brain-computer interface driven by speech, Biomedical Physics & Engineering Express 2 (6) (2016) 065009.
[6] V. Parekh, R. Subramanian, D. Roy, C. Jawahar, An eeg-based image annotation system, in: Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16–19, 2017, Revised Selected Papers 6, Springer, 2018, pp. 303–313.
[7] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, M. Shah, Brain2image: Converting brain signals into images, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 1809–1817.
[8] D. Nemrodov, M. Niemeier, A. Patel, A. Nestor, The neural dynamics of facial identity processing: Insights from eeg-based pattern analysis and image reconstruction, eNeuro (2018). ENEURO–0358
[9] P. Bashivan, I. Rish, M. Yeasin, N. Codella, Learning representations from eeg with deep recurrent-convolutional neural networks, arXiv:1511.06448 (2015).
[10] S. Opałka, D. Szajerman, B. Stasiak, A. Wojciechowski, Effective bci mental task classification with multichannel recurrent neural networks.
[11] M. Ilyas, P. Saad, M. Ahmad, A. Ghani, Classification of eeg signals for brain–computer interface applications: Performance comparison, in: 2016 International Conference on Robotics, Automation and Sciences (ICORAS), IEEE, 2016, pp. 1–4.
[12] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Hum Brain Mapp 38 (11) (2017) 5391–5420.
[13] R.M. Demirer, M.S. Ozerdem, C. Bayrak, Classification of imaginary movements in ecog with a hybrid approach based on multi-dimensional hilbert-svm solution, J. Neurosci. Methods 178 (1) (2009) 214–218.
[14] S.D. Stavisky, P. Rezaii, F.R. Willett, L.R. Hochberg, K.V. Shenoy, J.M. Henderson, Decoding speech from intracortical multielectrode arrays in dorsal arm/hand areas of human motor cortex, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 93–97.
[15] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6809–6817.
[16] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, M. Shah, Decoding brain representations by multimodal learning of neural activity and visual features, arXiv:1810.10974 (2018).
[17] A.X. Stewart, A. Nuthmann, G. Sanguinetti, Single-trial classification of eeg in a visual object task using ica and machine learning, J. Neurosci. Methods 228 (2014) 1–14.
[18] I. Simanova, M. Van Gerven, R. Oostenveld, P. Hagoort, Identifying object categories from event-related eeg: toward decoding of conceptual representations, PLoS ONE 5 (12) (2010) e14465.
[19] B. Kaneshiro, M.P. Guimaraes, H.-S. Kim, A.M. Norcia, P. Suppes, A representational similarity analysis of the dynamics of object processing using single-trial eeg classification, PLoS ONE 10 (8) (2015) e0135697.
[20] R. Li, J.S. Johansen, H. Ahmed, T.V. Ilyevsky, R.B. Wilbur, H.M. Bharadwaj, J.M. Siskind, Training on the test set? an analysis of spampinato et al. [arxiv: 1609.00344] (2018) arXiv:1812.07697.
[21] H.H. Jasper, The ten-twenty electrode system of the international federation, Electroencephalogr. Clin. Neurophysiol. 10 (1958) 370–375.
[22] E. Gysels, P. Renevey, P. Celka, Svm-based recursive feature elimination to compare phase synchronization computed from broadband and narrowband eeg signals in brain–computer interfaces, Signal Processing 85 (11) (2005) 2178–2189.
[23] J.-K. Kamarainen, Gabor features in image analysis, in: Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference on, IEEE, 2012, pp. 13–14.
[24] M. Chirimuuta, Explanation in computational neuroscience: causal and non–causal, Br J Philos Sci 69 (3) (2017) 849–880.
[25] S. Marčelja, Mathematical description of the responses of simple cortical cells, JOSA 70 (11) (1980) 1297–1300.
[26] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, JOSA A 2 (7) (1985) 1160–1169.
[27] X. Chen, F. Han, M.-m. Poo, Y. Dan, Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (v1), Proceedings of the National Academy of Sciences 104 (48) (2007) 19120–19125.
[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[29] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[30] C.M. Bishop, Pattern recognition and machine learning (information science and statistics), Springer-Verlag, Berlin, Heidelberg, 2006.
[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[32] R.S. Snell, Clinical neuroanatomy, Lippincott Williams & Wilkins, 2010.
[33] E.N. Marieb, K. Hoehn, Human anatomy & physiology, Pearson Education, 2007.
[34] G. Rizzolatti, L. Craighero, The mirror-neuron system, Annu. Rev. Neurosci. 27 (2004) 169–192.

Nicolae Cudlenco received his Bachelor of Computer Science (2016) and M.Sc. degree in Artificial Intelligence (2018) at University POLITEHNICA of Bucharest. Currently he is continuing his PhD in Computer Vision and Machine Learning at the School of Advanced Studies of the Romanian Academy. His current research interests lie at the intersection of Deep Learning, Computer Vision and eHealth systems.

Nirvana Popescu is full professor at University POLITEHNICA of Bucharest (UPB), Computer Science Department, since 2014. She received her Bachelor of Engineering in 1998 at UPB, Computer Science Department, followed by M.Sc at the same department. In 2003, she became PhD in Computer Science with a thesis called "Self-organizing intelligent fuzzy systems", at UPB. Meanwhile she worked and studied as PhD guest student at the University of Bielefeld, Germany. Her main research interests are neural networks, intelligent systems, fuzzy logic and control, eHealth systems, cognitive and autonomous robots, reconfigurable computers. At UPB, she is the leader of the Laboratory for "Reconfigurable high-confidence medical devices" and she coordinated research projects in the field of Medical Rehabilitation Robotics and Medical Wearable Sensors.

Marius Leordeanu is Associate Professor in Computer Science University "Politehnica of Bucharest and Senior Researcher at the Institute of Mathematics of the Romanian Academy. Marius leads several research groups in computer vision and robotics and has active R&D collaborations in AI with top IT companies in Romania. Marius has a PhD in Robotics from Carnegie Mellon University (2009) and Bachelor Degrees in Mathematics and Computer Science from the City University of New York (2003). He publishes in top computer vision and machine learning forums on topics such as recursive space-time graph neural networks, unsupervised visual learning in space and time, vision for drones and autonomous vehicles, semantic segmentation in video and vision to language translation. In 2014 Dr. Leordeanu was awarded the "Grigore Moisil" Prize in Mathematics, the top prize given by the Romanian Academy for his work on unsupervised learning for graph matching. He is actively involved in the growing of the AI community in Easter Europe (co-organizer of EEML and SSIMA international AI summer schools) and serves as Area Chair for top conferences (ICCV, CVPR and ECCV) and as Area Editor for high quality journals (CVIU, MVA).

Please cite this article as: N. Cudlenco, N. Popescu and M. Leordeanu, Reading into the mind's eye: Boosting automatic visual recognition with EEG signals, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.12.076